nestor

Anforderungen von e-Science und Grid-Technologie an die Archivierung wissenschaftlicher Daten

Jens Klump

GeoForschungsZentrum Potsdam

nestor-materialien 9

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

Anforderungen von e-Science und Grid-Technologie an die Archivierung wissenschaftlicher Daten

Jens Klump
GeoForschungsZentrum Potsdam

nestor-materialien 9

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Herausgegeben von

nestor - Kompetenznetzwerk Langzeitarchivierung und
Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland

nestor - Network of Expertise in Long-Term Storage of Digital Resources

http://www.langzeitarchivierung.de

Projektpartner:

Bayerische Staatsbibliothek, München
Bundesarchiv
Deutsche Nationalbibliothek (Projektleitung)
FernUniversität in Hagen
Humboldt-Universität zu Berlin - Computer- und Medienservice / Universitätsbibliothek
Institut für Museumsforschung, Berlin
Niedersächsische Staats- und Universitätsbibliothek, Göttingen

Betreuer für diese Veröffentlichung:

FernUniversität in Hagen
Prof. Dr.-Ing. Matthias L. Hemmje,
Dr. Dominic Heutelbeck,
Dr. Claus-Peter Klas,
Holger Brocks

URN: urn:nbn:de:0008-2008040103
[http://nbn-resolving.de/urn/resolver.pl?urn= urn:nbn:de:0008-2008040103]

Grid-Technologie und Langzeitarchivierung in nestor

Die modernen Informationstechnologien haben in allen Lebensbereichen starke Veränderungen bewirkt. Besonders stark beeinflusst sind die Wissenschaften, die auch eine treibende Kraft dieser Entwicklungen sind und immer größere Anforderungen an Rechner, Speicher und IT-Werkzeuge stellen. In neuen Experimenten der Teilchenphysik werden kaum bewältigbare Datenmengen für Tausende von Wissenschaftlern produziert, Klimaforscher berechnen immer detailliertere Modelle des Systems Erde, und die Geisteswissenschaften beginnen riesige digitale Sammlungen von Kulturgütern mit Rechnern zu analysieren.
Die Grid-Technologie zur Aufteilung der Aufgaben auf viele verteilte IT-Ressourcen ist ein Mittel, um den Herausforderungen dieser neuen, als e-Science bezeichneten wissenschaftlichen Arbeitsweise gerecht zu werden.

nestor und Wissenschaftler weltweit haben immer wieder darauf hingewiesen, dass mit der Zunahme der Bedeutung digitaler Daten auch die Notwendigkeit wächst, ihre langfristige Nutzbarkeit zu sichern. Bei der Grid-Technologie ergibt sich die chancenreiche Situation, dass nicht nur wertvolle und zu erhaltende Daten produziert werden, sondern auch Mittel bereit gestellt werden, die für die Herausforderung der Langzeitarchivierung großer und komplexer Datenmengen nutzbar sein können. Die klassischen Gedächtnisorganisationen - wie Bibliotheken, Archive und Museen - und die neuen Gedächtnisorganisationen - wie Daten- und Rechenzentren - können wechselseitig voneinander profitieren.

Um dieses Potenzial auszuloten, hat nestor in seiner zweiten Projektphase eine Arbeitsgruppe mit Fachleuten aus klassischen Gedächtnisinstitutionen und aus e-Science- und grid-engagierten Institutionen initiiert und drei Expertisen in Auftrag gegeben. Diese Expertisen untersuchen den Ist-Stand und die Anforderungen und Ziele für das Zusammenspiel von e-Science-/Grid-Technologie und Langzeitarchivierung unter drei Gesichtspunkten:
Welche Anforderungen gibt es für die Archivierung von Forschungsdaten?
Was sind die möglichen Synergien, die angestrebt werden sollten?
Und auf welche Standards können weitere Arbeiten in diesen Bereich aufgebaut werden und welche sind gegebenenfalls noch zu entwickeln?

Neben der Untersuchung des Standes der Technik, sind einige Projekte der deutschen Grid-Initiative D-Grid befragt worden. nestor wird in seiner Grid-/eScience-Arbeitsgruppe die Ergebnisse der Expertisen aufnehmen und versuchen, eine Landkarte für die weiteren Entwicklungsperspektiven zu zeichnen.

e-Science-/Grid-Technologie und Langzeitarchivierung sind relativ neue Forschungsbereiche, die sich sehr schnell entwickeln. Einzelne Fragen, die von nestor Mitte 2006 formuliert wurden, als die ersten Projekte der deutschen Grid-Initiative D-Grid gerade gestartet waren, stellen sich heute, wo bald schon die dritte Generation von D-Grid-Projekten beginnt, unter den veränderten Bedingungen möglicherweise anders dar. Die Expertisen müssen daher auch vor ihrem Entstehungshintergrund betrachtet werden. Derzeit liefern sie eine Beschreibung sinnvoller und notwendiger Entwicklungen. Wenn sie in naher Zukunft „veralten“, weil sie zur erfolgreichen Zusammenarbeit von e-Science/Grid und Langzeitarchivierung beigetragen haben, dann haben sie ihren Sinn erfüllt.

Anforderungen von e-Science und Grid-Technologie an die Archivierung wissenschaftlicher Daten

Jens Klump
GeoForschungsZentrum Potsdam

Inhaltsverzeichnis

Zusammenfassung 3

1 Einleitung und Stand der Dinge 5
1.1 Zielsetzung der Studie 6
1.2 Begriffsdefinitionen 7
1.3 Nationale und internationale Aktivitäten 10

2 Ergebnisse der Studie 12
2.1 Herausforderung Archivtechnologie 14
2.2 Herausforderung Metadaten 22
2.3 Herausforderung Semantic Web 26
2.4 Herausforderungen Zugang zu Daten und Rechteverwaltung 30
2.5 Herausforderung Organisation und Nachhaltigkeit 33

3 Handlungsempfehlungen 36
3.1 Technik 36
3.2 Metadaten 37
3.3 Semantic Grid 37
3.4 Rechteverwaltung 37
3.5 Organisation von Virtuellen Organisationen 38

Danksagung 39

Literatur 40

Anhang – Fragebogen
„Anforderungen von eScience und Grid-Technologie an die Archivierung wissenschaftlicher Rohdaten“ 43

Zusammenfassung

Die enorm großen Datenmengen, die in Grid-Projekten erzeugt und verarbeitet werden und die hohe Komplexität von Daten aus eScience-Projekten lassen vermuten, dass aus diesen Projekttypen neuartige Anforderungen an die digitale Langzeitarchivierung erwachsen. Umgekehrt besteht die Möglichkeit, dass aus der Grid-Technologie oder aus den semantischen Werkzeugen der eScience-Projekte neue Methoden entstehen, die der digitalen Langzeitarchivierung eingesetzt werden können.

Die Expertise „Anforderungen von eScience und Grid-Technologie an die Archivierung wissenschaftlicher Daten“ untersucht aus technologischer wie organisatorisch-strategischer Perspektive, ob existierende e-Science-Infrastrukturen in Rohdatenproduzierenden Communities den Anforderungen zur Langzeitarchivierung gerecht werden, und ob die Erfahrungen der Communities im Bereich der Grid-Technologien auf Organisationen und Systeme zur digitalen Langzeitarchivierung übertragen werden können.

Im Umgang mit den Anforderungen der digitalen Langzeitarchivierung werden zwischen den befragten Projekten große Unterschiede sichtbar. Die erreichten Ergebnisse, aber auch die vorgefundenen Defizite, werden in dieser Studie vorgestellt und diskutiert und Problemfelder analysiert. Die Handlungsempfehlungen sind aus den Ergebnissen der Befragung und der Analyse der Herausforderungen, mit denen Grid- und eScience-Projekte in der digitalen Langzeitarchivierung konfrontiert werden, abgeleitet.

Executive Summary

The enormous amounts of data from Grid projects and the complexity of data from e-science projects suggest that these new types of projects also have new requirements towards long-term archiving of data. On the other hand, Grid technology and semantic tools emerging from e-science might provide us with new methods that may be useful in long-term digital preservation.

The study “Requirements of e-science and Grid projects towards long-term archiving of scientific and scholarly data” investigates from a technological and from a management perspective whether existing infrastructures in data producing research e-science and Grid communities meet the requirements of long-term digital preservation. The study also investigates, whether technologies and best practices from e-science and Grid project can be transferred to organisations and systems in the field of long-term digital preservation.

The interviews conducted as part of this study showed considerable differences between projects in the way they approached long-term digital preservation of data. Their achievements –but also their deficits– are analysed and discussed. The recommendations given in this study are derived from this analysis and discussion.

Recommendations:

1. **Technological Requirements**
   - Development of a test-bed for the application of Grid services in long-term digital preservation.
   - Identification of standards in the Grid environment that are relevant to long-term digital preservation.

- Research in to solutions for the re-use of digital objects from obsolete software and hardware platforms (virtual machines, emulation vs. migration).

- Development of criteria for the evaluation of file formats and their fitness for long-term digital preservation.

- Communication of best-practices in long-term digital preservation to e-science and Grid projects to improve data stewardship.

2. Metadata

- Communication of best-practices in metadata generation and processing to improve metadata practice in e-science and Grid projects.

- Development of tools for the automatic of semi-automatic generation of metadata and integration of these tools into the scientific workflow. Currently, the emphasis is on discovery metadata. Future research should broaden the scope to metadata describing data provenance and lineage and the encoding of explicit and implicit knowledge.

3. Semantic Grid

- Transfer of semantic web technologies from e-science to Grid projects to improve semantic interoperability between community Grids and to improve the re-use of already existing data.

- Expand the use of global unique identifiers for the unambiguous identification of datasets. More work is needed to uniquely reference small subsets of very large datasets.

- Development of preview formats for large and multidimensional datasets.

4. Digital Rights Management

- Evaluation of currently deployed mechanisms for authentication, authorisation and access control for secure and continuous operation over very long periods of time. In particular, it has to investigated whether currently deployed technologies in the Grid environment allow a secure transfer of the policies governing authentication and authorisation infrastructures to future security technologies.

- Transfer of Identity 2.0 concepts, like Identity Credential Services, to authentication and authorisation technologies in addition to those currently used in the Grid environment.

5. Management of Virtual Organisation

- Research into improved management models for Virtual Organisations and incentives to improve long-term digital preservation of data from e-science and Grid projects.

- Development of concepts for education, training and professional development of long-term digital preservation of research data in the e-science and Grid environment.

- Communication of best-practice examples in long-term digital preservation to e-science and Grid projects.

1 Einleitung und Stand der Dinge

In der wissenschaftlichen Forschung produzierte Daten sind in vielen Sektoren von zentraler Bedeutung, denn neben Theorie und Experiment hat sich in den letzten Jahrzehnten die rechnergestützte quantitative Analyse und Modellierung als „dritte Säule“ wissenschaftlicher Tätigkeit etabliert. Diesen Daten stammen nicht allein aus Messungen oder Experimenten der Naturwissenschaften, sondern auch aus Quellen in den Geistes- und Sozialwissenschaften, zum Beispiel aus soziologische Panel-Befragungen oder aus der linguistische Analyse von Texten. (Nature Redaktion, 2006). Sowohl öffentliche Institutionen wie auch Wirtschaftsunternehmen investieren erhebliche Mittel in die Produktion von Rohdaten und das jährlich produzierte Volumen an Rohdaten steigt stetig an (Kroker, 2006). Damit gewinnt auch die Forderung nach deren Verfügbarkeit zur möglichen Nachprüfung von wissenschaftlichen Ergebnissen und zur Wiederverwendung große Bedeutung (Klump et al., 2006). Voraussetzung für diese Art von Wandel in der Forschung ist die digitale Langzeitarchivierung von Forschungsdaten.

Die enorm großen Datenmengen, die in Grid-Projekten erzeugt und verarbeitet werden und die hohe Komplexität von Daten aus eScience-Projekten lassen jedoch vermuten, dass aus diesen Projekttypen neuartige Anforderungen an die digitale Langzeitarchivierung erwachsen (Hey und Trefethen, 2003a). Gerade wegen dieser extremen Anforderungen an Prozessierungs- und Speicherressourcen und zusätzlichen Managementvorkehrungen durch Virtualisierung der Ressourcen sind Communities, die große Datenmengen erzeugen oder verarbeiten, in der Anwendung von Grid-Technologien vergleichsweise weit fortgeschritten. Astrophysik, Klimatologie, biomedizinische Forschung, und andere Communities mit rechenintensiven Verfahren der Datenverarbeitung wenden bereits seit einiger Zeit Grid-Technologien an.

Umgekehrt besteht die Möglichkeit, dass aus der Grid-Technologie oder aus den semantischen Werkzeugen der eScience-Projekte neue Methoden entstehen, die der digitalen Langzeitarchivierung eingesetzt werden können. Sofern der Hinweis auf solche Synergien von den befragten Projekten selbst genannt wurden, werden sie auch in dieser Studie dargestellt. In zwei weiteren nestor-Studien werden Synergien zwischen Grid-Technologie und digitaler Langzeitarchivierung (Schiffmann, in prep.), sowie Standardisierungsbedarf beim Einsatz von Grid-Technologie in der digitalen Langzeitarchivierung (Borghoff und Rödig, in prep.) detailliert untersucht.

Anders als in Großbritannien standen in Deutschland bei den Themen eScience und Grid bisher technologische Überlegungen im Vordergrund. Die Herausforderungen, die aus dem Umgang mit Daten erwachsen, wurden bislang wenig diskutiert (Schroeder et al., 2007). Die Expertise „Anforderungen von eScience und Grid-Technologie an die Archivierung wissenschaftlicher Daten“ soll sowohl aus technologischer wie organisatorisch-strategischer Perspektive prüfen, ob existierende e-Science-Infrastrukturen in Rohdatenproduzierenden Communities den Anforderungen zur Langzeitarchivierung (länger als den in der Community üblichen Aufbewahrungsdauern von etwa 10 Jahren) gerecht werden können, und ob die Erfahrungen der Communities im Bereich der Grid-Technologien auf Organisationen und Systeme zur Langzeitarchivierung übertragen werden können.

1.1 Zielsetzung der Studie

Aus diesen beiden Näherungsansätzen - Grid-Technologien für die Langzeitarchivierung, und Langzeitarchivierung in die eScience-Community - ergeben sich weitere wichtige Fragestellungen, unter anderem:

- Inwiefern können Grid-Technologien die Verfügbarkeit von Rohdaten und deren Stabilität erhöhen?
- Welche existierenden eScience-Konzepte aus den Communities, die Rohdaten produzieren, sind aus der Perspektive der Archivierung besonders interessant (im positiven wie im negativen Sinn), wenn sie auch nicht per se für Archivierungszwecke etabliert worden sind?
- Können diese Konzepte auf andere Communities und Umgebungen übertragen werden?
- Wo genügen diese Ansätze und wo greifen sie aus Sicht der Langzeitarchivierung zu kurz?
- Entstehen durch die Anwendung von Grid-Technologien spezielle Anforderungen an Metadaten oder andere Systemkomponenten? Gibt es dazu Community übergreifende Ansätze und Standards unter den Rohdaten erzeugenden Organisationen und welche Auswirkung könnten diese auf Metadaten zur Archivierung wie PREMIS[1], den Archivstandard OAIS[2] oder andere verbreitete Archivierungskonzepte haben?
- Gibt es sektorspezifische Anforderungen an die Archivierung, die eine besondere Herausforderung an Grid-Technologien darstellen (beispielsweise im Bereich der Medizin, in dem besonders sensible Daten archiviert werden müssen und Datenschutz von hoher Bedeutung ist)?
- Welche Auswirkungen hat das verteilte Management von Daten auf der Strategieebene? Die verteilten Speicherressourcen müssen auch von jemandem zur Verfügung gestellt und gewartet werden.
- Welche organisatorischen Konstellationen zur Unterhaltung einer Grid-Infrastruktur existieren bereits in den Rohdaten produzierenden Communities und könnten diese Muster auf entsprechende Infrastrukturen für Archivierungszwecke übertragen werden?

Bereits in der ersten Förderphase des Projekts nestor untersuchten Severins und Hilf (Severiens und Hilf, 2006a; Severiens und Hilf, 2006b), welche Anforderungen an die digitale Langzeitarchivierung wissenschaftlicher Daten gestellt werden. Diese umfangreichen Vorarbeiten sollen hier nicht dupliziert werden. In vielen Aspekten unterscheidet sich die digitale Langzeitarchivierung von Forschungsdaten aus eScience- und Grid-Projekten auch nicht wesentlich von anderen Datenproduzierenden Forschungsprojekten oder von den allgemeinen Grundsätzen der digitalen Langzeitarchivierung. Deshalb sollen hier auch diese Arbeiten aus dem nestor-Projekt und anderen Projekten zur digitalen Langzeitarchivierung nicht dupliziert werden. Der Leser wird daher an Stellen, die über den Rahmen dieser Studie hinaus gehen, auf die weiterführende Literatur verwiesen.

[1] PREMIS: PREservation Metadata: Implementation Strategies Working Group, eine Arbeitsgruppe der Library of Congress und der Firma OCLC zur Standardisierung von Metadaten zur digitalen Langzeitarchivierung (http://www.loc.gov/standards/premis/).

[2] OAIS: Open Archival Information System - Archivmodell (ISO 14721). Es beschreibt ein Referenzmodell, in dem Menschen und Systeme zusammenwirken, um digitale Objekte zu erhalten und definierten Nutzergruppen zugänglich machen

In dieser Studie wird die Frage nach neuen Lösungsansätzen aus der Grid-Technologie für die digitale Langzeitarchivierung von Forschungsdaten nur unter der Fragestellung der daraus ableitbaren Nutzerbedürfnisse diskutiert. Eine separate nestor-Expertise ist allein diesem Thema gewidmet (Schiffmann, in prep.).

1.2 Begriffsdefinitionen

In der Diskussion zur digitalen Langzeitarchivierung von Primärdaten aus eScience- und Grid-Projekten ist nicht immer klar, wie die genannten Begriffe zu verstehen sind. Aus diesem Grund sollen in den nachfolgenden Unterabschnitten die Begriffe „Primärdaten“, „eScience“, „Grid“ und „Langzeitarchivierung“ für diese Studie definiert werden.

1.2.1 Primärdaten

Der Begriff „Primärdaten“ sorgt immer wieder für Diskussion, denn die Definition des Begriffs ist sehr von der eigenen Rolle in der wissenschaftlichen Wertschöpfungskette geprägt. Für den einen sind „Primärdaten“ der Datenstrom aus einem Gerät, z.B. einem Satelliten. In der Fernerkundung werden diese Daten „Level 0“ Produkte genannt. Für einen anderen sind „Primärdaten“ zur Nachnutzung aufbereiteten Daten, ohne weiterführende Prozessierungsschritte. Andere wiederum differenzieren nicht nach Grad der Verarbeitung sondern Betrachten nur die Daten, die Grundlage einer wissenschaftlichen Veröffentlichung waren. Der US National Research Council, der die National Science Foundation wissenschaftlich und politisch berät, definiert Primärdaten als „facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors” und „Public Data“ als „data that are generated through research within government organizations, or by academic or other not-for-profit entities, as well as public data used for research purposes, but not necessarily produced primarily for research (e.g., geographic or meteorological data, or socioeconomic statistics produced by or for government organizations).“ (Uhlir und Schröder, 2007).

Durch eine Reihe von Aufsehen erregenden Wissenschaftsskandalen in den neunziger Jahren des 20. Jahrhunderts sah sich die Deutsche Forschungsgemeinschaft (DFG) gezwungen, „Regeln für einen gute wissenschaftliche Praxis“ auszusprechen (DFG, 1998), die in vergleichbarer Form auch von anderen Wissenschaftsorganisationen übernommen wurden. Für den Umgang mit Daten bezieht sich die DFG auf Daten, die Grundlage einer wissenschaftlichen Veröffentlichung waren. Sie verlangt von ihren Zuwendungsempfängern, dass diese Daten für mindestens zehn Jahre auf geeigneten Datenträgern sicher aufbewahrt werden müssen (DFG, 1998, Empfehlung 7). Für die einzelnen Disziplinen ist der Umgang mit Daten im einzelnen zu klären, um eine angemessene Lösung zu finden (DFG, 1998, Empfehlung 1). Diese Policy dient jedoch in erster Linie einer Art Beweissicherung, über Zugang zu den Daten und ihre Nachnutzbarkeit sagen die Empfehlungen nichts aus.

Auf Grund der enormen Summen, die jährlich für die Erhebung wissenschaftlicher Daten ausgegeben werden beschäftigt sich die Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) bereits seit einigen Jahren mit der Frage, wie mit Daten aus öffentlich geförderter Forschung umgegangen werden soll. Auf dem Treffen der Forschungsminister im Januar 2004 wurde beschlossen, dass der Zugang zu Daten aus öffentlich geförderter Forschung verbessert werden muss (OECD, 2004). Mit diesem Mandat im Hintergrund befragte die OECD die Wissenschaftsorganisationen ihrer Mitgliedsländer zu deren Umgang mit Forschungsdaten. Aus dem Ergebnissen der Befragung wurde eine Studie verfasst und im Dezember 2006 verabschiedete der Rat der OECD eine „Empfehlung betreffend den Zugang zu Forschungsdaten aus öffentlicher Förderung“ (OECD, 2006). Diese Empfehlung ist

bindend und muss von den Mitgliedsstaaten der OECD in nationale Gesetzgebung umgesetzt werden, die Umsetzung wird von der OECD beobachtet. In Abschnitt M der Empfehlung wird vorgeschlagen, dass schon bei der Planung von Projekten eine nachhaltige, langfristige Archivierung der Daten berücksichtigt wird.

Parallel dazu, und mit mehr Aufsehen in der Öffentlichkeit, wurde im Oktober 2003 von den Wissenschaftsorganisationen die „Berliner Erklärung über den offenen Zugang zu wissenschaftlichem Wissen" veröffentlicht (Berliner Erklärung, 2003), deren Schwerpunkt auf dem Zugang zu wissenschaftlicher Literatur für Forschung und Lehre liegt. In ihre Definition des offenen Zugangs bezieht die „Berliner Erklärung" auch Daten und Metadaten mit ein. Die Langzeitarchivierung ist hier ein Mittel zum Zweck, dass den offenen Zugang zu wissenschaftlichem Wissen über das Internet auf Dauer ermöglichen soll.

Wenngleich es einige Policies gibt, die den Zugang zu Daten ermöglichen sollen, so hat sich erst recht spät die Erkenntnis durchgesetzt, dass die digitale Langzeitarchivierung von Forschungsdaten eine Grundvoraussetzung des offenen Zugangs ist. Eine Umfangreiche Studie wurde dazu bereits in der ersten Förderphase des Projekts nestor erstellt (Severiens und Hilf, 2006b). Eine ähnliche Studie wurde auch für das britischen Joint Information Systems Committee (JISC) erstellt (Lord und Macdonald, 2003) und das Thema in einer weiteren Studie vertieft (Lyon, 2007).

1.2.2 eScience

> *„eScience ist die globale Zusammenarbeit in Schlüsselgebieten der Forschung und die nächste Generation Werkzeuge, um diese Art von Forschung zu ermöglichen"*
>
> Taylor in (Hey und Trefethen, 2003a).

Die oben zitierte Definition von eScience nach Taylor ist charakteristisch für das in Großbritannien entwickelte Verständnis von eScience. In Deutschland war eine parallele Entwicklung von eScience- und Grid-Projekten zu beobachten, wobei die eScience-Projekte kaum die von den Grid-Projekten angebotenen Technologien nutzten, andererseits die angebotenen Grid-Dienste als wenig nutzerfreundlich angesehen wurden (Schroeder et al., 2007). Aus dieser Vorgeschichte heraus ist für die vom Bundesministerium für Bildung und Forschung (BMBF) geförderten eScience-Projekte die hohe semantische Komplexität charakteristisch, mit der Daten, Dokumenten und interaktiven Werkzeugen zu deren Bearbeitung miteinander verknüpft sind, die verarbeiten Datenmengen bleiben jedoch vergleichsweise gering. Andererseits gibt es unter den vom BMBF geförderten Grid-Projekten durchaus solche, die Objekte von hoher semantischer Komplexität verwalten und damit auch Charakteristika von eScience-Projekten aufweisen.

Im Allgemeinen werden Semantic Web Technologien wie RDF[3], OWL[4] oder darauf aufbauend SKOS[5] eingesetzt um die Beziehungen zwischen den Objekten zu beschreiben. In einzelnen Projekten gibt es Bestrebungen, auch Beziehungen zwischen Datenobjekten und Objekten der physischen Welt mit zu beschreiben und zu verwalten, Ansätze zum sog. „Internet der Dinge".

[3] Resource Description Framework, RDF, http://www.w3.org/RDF/
[4] Web Ontology Language, OWL, http://www.w3.org/2004/OWL/
[5] Simple Knowledge Organisation System, SKOS, http://www.w3.org/2004/02/skos/

1.2.3 Grid

„Das Grid stellt standardisierte Schnittstellen zu verteilten Rechen-, Speicher- und Bandbreitenressourcen einer heterogenen Infrastruktur sowie komplexen Dienstleistungen bereit.“

(Berman et al., 2003).

Auf Grund des hohen Bedarfs an Rechen-, Speicher- und Bandbreitenressourcen wurde die Entwicklung der Grid-Technologie als Forschungswerkzeug bisher in erster Linie von Projekten aus den Naturwissenschaften, z.B. Hochenergiephysik, Astrophysik oder Bioinformatik, vorangetrieben, aber auch für andere rechenintensive Fragestellung, wie z.B. die linguistische Analyse von Texten, wird Grid-Technologie angewendet. Unter den vom BMBF geförderten Grid-Projekten befindet sich nur ein einziges Projekt aus den Geisteswissenschaften. Diese Projekte produzieren Datenmengen, die teilweise weit über die bisher in diesen Disziplinen üblichen Datenmengen hinaus gehen. Die in Grid-Projekten erzeugten Datenmengen langfristig zu sichern und für eine wissenschaftliche Nachnutzung verfügbar zu machen stellt eine bisher nicht gekannte Herausforderung dar (Hey und Trefethen, 2003a).

Neben neuen Herausforderungen an die digitale Langzeitarchivierung bietet Grid-Technologie jedoch auch Werkzeuge an, die für die digitale Langzeitarchivierung durchaus nützlich sein könnten. Denkbar sind z.B. Synergien durch Nutzung des Daten-Grid oder durch das Auslagern ressourcenintensiver Archivprozesse, z.B. im Archiv-Ingest oder für Formatkonversionen, indem Prozesse der Langzeitarchivierung an externe Dienste ausgelagert werden (Hitchcock et al., 2007). Diese potenziellen Synergien zwischen Grid-Technologie und digitaler Langzeitarchivierung werden in einer separaten nestor-Expertise beleuchtet (Schiffmann, in prep.). Um zu erfahren, ob die bereits laufenden Projekte sich dieser Synergiepotenziale bewusst sind, wurden sie jedoch bereits in dieser Studie dazu befragt.

1.2.4 Langzeitarchivierung

“Digital information lasts forever — or five years, whichever comes first.”

(Rothenberg, 1997)

Langzeitarchivierung von Daten aus Forschungs- und Entwicklungsprojekten bezeichnet die nachnutzbare und vertrauenswürdige Archivierung von Daten. Die Dauer der Archivierung über das Ende des Projektes hinaus wird durch eine Policy zur digitalen Langzeitarchivierung oder durch den gesetzlichen Rahmen des Projekts bestimmt.

Langzeitarchivierung unterscheidet sich von Datenspeicherung (Storage) und Datensicherung (Backup) dadurch, dass die Stabilität, Integrität und Nachnutzbarkeit der Daten über einen langen Zeitraum angestrebt wird. Die der Langzeitarchivierung zu Grunde liegende Strategie berücksichtigt dabei technische Entwicklungen, wie z.B. eventuell notwendige Auffrischung oder Änderungen der Speichermedien, Formate, Abspielumgebungen, und Änderungen der organisatorischen Rahmenbedingungen Archivs.

Das Archivmodell nach ISO 14721 (Open Archival Information System - OAIS) beschreibt ein Referenzmodell, in dem Menschen und Systeme zusammenwirken, um digitale Objekte zu erhalten und definierten Nutzergruppen zugänglich machen (OAIS, 2002). In separaten nestor-Expertisen wird untersucht, in welche Funktionen der Archivierungsprozesse Grid-Technologien eingesetzt werden können und welcher Bedarf zur Standardisierung bei der Nutzung des Grid zur digitalen Langzeitarchivierung besteht.

Bei sehr großen Datenmengen müssen Auswahlkriterien definiert werden, nach denen Forschungsdaten langzeitarchiviert werden oder nur für eine begrenzte Zeit zur Verfügung stehen. Die Auswahlkriterien des British Atmospheric Data Centre (BADC) sind eines der wenigen Beispiele in denen bisher Auswahlkriterien für die Langzeitarchivierung von Daten als Regelwerk formuliert und niedergeschrieben wurden (Lyon, 2007).

Ein wichtiges Element des OAIS-Referenzmodells ist die Nutzergruppe, für die das System betrieben wird. Die Ausgestaltung der Prozesse orientiert sich deshalb an der Nachnutzbarkeit der archivierten Objekte durch die vorgesehene Nutzergruppe. Besonders wichtig für die Nachnutzbarkeit sind daher auch Datenformate und Metadatenschemata, denn erst wenn diese durch Nutzer in der Zukunft gelesen und ausgewertet werden können, ist eine Nachnutzung der archivierten Objekte möglich.

1.3 Nationale und internationale Aktivitäten

Der Bedarf an Forschung zur Digitale Langzeitarchivierung von Forschungsdaten ist bekannt und wird in einer Reihe von deutschen, europäischen und internationalen Projekten bearbeitet uns soll nicht unerwähnt bleiben. Eine Auswahl von ihnen soll hier kurz dargestellt werden.

Kompetenznetzwerk Langzeitarchivierung (nestor)
Das Kompetenznetzwerk nestor[6] verfolgt das Ziel, die digitalen Ressourcen in Deutschland zu sichern und verfügbar zu machen sowie mit anderen Netzwerken und Entscheidungsträgern national und international zusammenzuarbeiten, um gemeinsam die digitale Wissensbasis langfristig zu bewahren. Die notwendigen Fachkompetenzen für den Aufgabenkomplex "Langzeitarchivierung digitaler Ressourcen" verteilen sich über ein breites Spektrum von Personen, die in vielen Institutionen, Organisationen und Wirtschaftsunternehmen tätig sind. Dieses Wissen soll durch nestor vernetzt und zugänglich gemacht werden. nestor wird vom BMBF gefördert.

Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen (kopal)
Digitale Dokumente langfristig zur Verfügung zu stellen, ist ein bislang ungelöstes Problem unserer Informationsgesellschaft. Mit der ansteigenden Zahl elektronischer Veröffentlichungen wächst die Notwendigkeit einer zuverlässigen Archivierung. Im Zuge der technischen Entwicklung werden immer neue digitale Dateiformate verwendet, die an spezielle Programme und damit an bestimmte Rechnertypen und Betriebssysteme gebunden sind. Ältere Daten sind so mit aktueller Soft- und Hardware oft nicht mehr nutzbar. Das Projekt kopal[7] widmet sich der Lösung dieser Problematik in Form eines kooperativ entwickelten und betriebenen Langzeitarchivs für digitale Daten.

Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (CASPAR)
Das Projekt Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (CASPAR)[8] wird von der Europäischen Kommission im Rahmen des sechsten Rahmenprogramms für Forschung und Technologie (FP6) gefördert. Das Projekt betreibt Forschung zur Umsetzung innovativer Lösungen für die digitale Langzeitarchivierung basierend auf dem OAIS Referenzmodell (ISO 14721:2002) (OAIS, 2002). Durch enge Zusammenarbeit mit den jeweiligen Communities sollen mehr und vielseitigere Systeme und Services für die digitale Langzeitarchivierung entstehen. Ein Ziel von CASPAR ist es, Grid-Dienste für digitale Bibliotheken zu erschließen.

[6] nestor: http://www.langzeitarchivierung.de/
[7] kopal: http://kopal.langzeitarchivierung.de/
[8] CASPAR: http://www.casparpreserves.eu/

Preservation and Long-term Access through NETworked Services (PLANETS)
Durch die technische Entwicklung sind heute aktuelle Dateiformate eines Tages obsolet und die Information, die in diesen Dateien enkodiert ist, kann nur noch unter großem Aufwand gelesen werden. Viel wertvolle und einmalige Information geht so verloren. Das Projekt PLANETS[9] entwickelt Werkzeuge für die Planung digitaler Langzeitarchivierung um durch Automatisierung und skalierbare Architekturen deren Kosten in einem planbaren Rahmen zu halten. Ein weiterer Aspekt ist die Charakterisierung der Formate digitaler Objekte um deren Erhaltung durch bessere Planung der Archivprozesse zu unterstützen. Neben der Verbreitung der Projektergebnisse sollen diese auch für eine kommerzielle Nachnutzung durch Dienstleister für digitale Langzeitarchivierung nutzbar gemacht werden.

Sustaining Heritage Access through Multivalent ArchiviNg (SHAMAN)
Im Projekt Sustaining Heritage Access through Multivalent ArchiviNg (SHAMAN) werden Entwicklungen aus den Bereichen der digitalen Bibliotheken, der Daten-Grids und der digitalen Langzeitarchive zusammen geführt, um eine vollständige technische Umgebung für die digitale Langzeitarchivierung aufzubauen. Durch den Einsatz von Daten-Grid Technologie sollen Archivprozesse automatisiert werden. Durch neue Werkzeuge für das Management digitaler Sammlungen sollen die Möglichkeiten des Daten-Grid erweitert werden. Das Konzept soll an drei sich gegenseitig ergänzenden Anwendungsfällen validiert werden. Technisch baut das Projekt vor allem auf den Storage Resource Broker und auf iRODS auf (siehe auch Abschnitt 2.4.1).

Digital Repository Infrastructure Vision for European Research (DRIVER)
Das EU-Projekt Digital Repository Infrastructure Vision for European Research (DRIVER)[10] hat sich zum Ziel gesetzt, wissenschaftliche Literatur, experimentelle und Beobachtungsdaten und andere digitale Objekte über internetbasierte Infrastrukturen zugänglich zu machen. Es soll ein ergänzendes Wissensnetzwerk zum europäischen Rechennetzwerk GEANT2 aufbauen. Im Fordergrund stehen bei DRIVER der Zugang zu Wissen über standardisierte, vertrauenswürdige und zuverlässige Schnittstellen, auf die neue Wertschöpfungsketten aufgebaut werden können.

Digital Perservation Europe (DPE)
Ähnlich wie das Projekt nestor auf nationaler Ebene vernetzt das EU-Projekt Digital Perservation Europe (DPE)[11] in Europa vorhandene Kompetenz in der digitalen Langzeitarchivierung. Durch die Aktivitäten des Projekts sollen nicht nur Netzwerke geknüpft werden, sondern auch die Forschung zur digitalen Langzeitarchivierung koordiniert werden und die Ergebnisse in die Praxis überführt werden.

Die hier nur kurz aufgezählten Projekte zeigen, dass das Problem der Langzeitarchivierung unseres digitalen Kulturgutes erkannt wurde und an Lösungswegen gearbeitet wird. Zusätzlicher Forschungsbedarf entsteht jedoch aus den spezifischen Anforderungen der Grid- und eScience-Projekte und hier liegt auch der Fokus dieser Studie. Allgemeine Aspekte der digitalen Langzeitarchivierung sollen daher hier nur so weit dargestellt und diskutiert werden, wie sie für die unmittelbare Fragestellung der Studie relevant sind. Für Einzelheiten der Projekte sei daher auf diese selbst und auf die Literatur verwiesen.

[9] PLANETS : http://www.planets-project.eu/
[10] DRIVER: http://www.driver-repository.eu/
[11] DPE: http://www.digitalpreservationeurope.eu/

2 Ergebnisse der Studie

Die Befragung der vom BMBF geförderten eScience- und Grid-Projekte hat gezeigt, dass die digitale Langzeitarchivierung von Daten aus eScience- und Grid-Projekten neue Fragen aufwirft. Dabei geht es nicht nur um technische Fragen der Archivierung großer und komplexer Datenbestände, sondern auch um Fragen der Organisation von wissenschaftlichen Arbeitabläufen und notwendiger Rahmenbedingungen für eine erfolgreiche Langzeitarchivierung digitaler Forschungsdaten.

Ohne Zweifel lassen sich punktuell Lösungen für die technischen Herausforderungen von Grid- und eScience-Projekten an die digitale Langzeitarchivierung von Forschungsdaten finden. So bietet z.B. die Grid-Technologie die Möglichkeit, Ressourcenintensiven Prozesse kurzzeitig auszulagern. In dieser Studie soll jedoch der gesamte Prozess der digitalen Langzeitarchivierung betrachtet werden, und wie dieser mit digitalen wissenschaftlichen Wertschöpfungsketten verbunden ist (Klump et al., 2007). Als Modell für die Darstellung von Prozessen in digitalen Langzeitarchiven hat sich das Reference Model for an Open Archival Information System (OAIS) nach ISO-Standard 14721:2003 bewährt (OAIS, 2002). Unter den in Abschnitt 2.1.5 genannten Best Practice Beispielen befinden sich Archive, die ihre Prozesse gemäß dem OAIS-Referenzmodell aufgebaut haben (Eastman et al., 2005; Lyon, 2007).

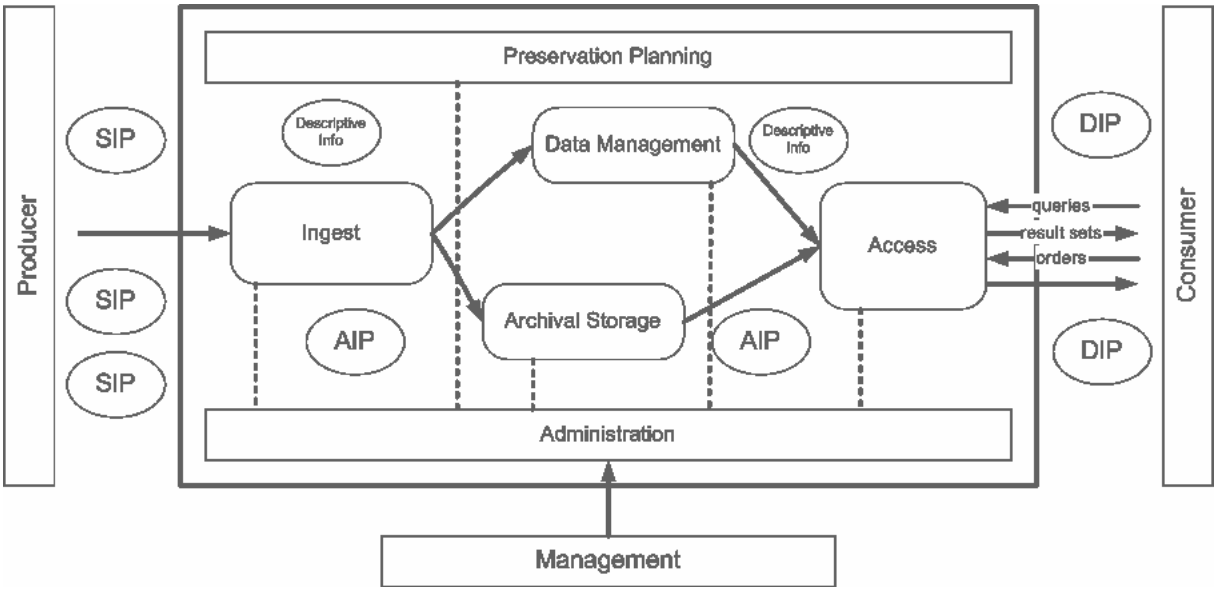

Abbildung 1: *Schematische Darstellung des OAIS-Referenzmodells aus Hitchcock et al. (2007). Vom Datenproduzenten gelieferte Übergabeprodukte (Submission Information Packages, SIP) werden in das Archiv aufgenommen (Ingest Prozesse) und in Archivobjekten (Archival Information Packages, AIP) abgelegt (Archival Storage Prozesse). Die Metadaten (Descriptive Info[rmation]) werden parallel zu den AIPs verwaltet (Data Management Prozesse). Daten-Konsumenten können nach archivierten Objekten suchen und erhalten Zugang zu ihnen über entsprechende Schnittstellen (Access Prozesse). Über diese Schnittstellen erhält der Konsument die angeforderten Nutzungsobjekte (Dissemination Information Packages, DIP).*

Als Ausgangspunkt der Expertise wurde untersucht, wie bereits bestehende Projekte mit diesen Herausforderungen umgehen und welche Lösungsansätze sie verfolgen. Um eine

möglichst enge Abstimmung der Expertise mit den Erfahrungen und Bedürfnissen der Anwender zu erreichen, wurden ein Fragebogen entwickelt, der als Gesprächsleitfaden für Interviews und Telefoninterviews diente oder auch schriftlich beantwortet wurde. Die Ergebnisse der Studie werden diskutiert und verschiedene Lösungsansätze erwogen, um daraus Handlungsempfehlungen für die digitale Langzeitarchivierung von Daten aus eScience- und Grid-Projekten abzuleiten.

Ebenso interessant wie die Auswertung der von den Projekten aufgeworfenen Fragen zur digitalen Langzeitarchivierung ist es zu Fragen, welche Themen nicht angesprochen wurden, sei es, weil die Anforderungen als lösbar betrachtet werden, oder – noch viel wichtiger – weil in Zukunft auftretende Probleme noch nicht wahrgenommen werden, z.B. weil sie über den Zeithorizont des Projekts hinaus gehen.

Als Zielgruppe wurden die im März 2007 vom BMBF geförderten Grid- und eScience-Projekte gewählt[12], da hier die Ansprechpartner relativ leicht zu identifizieren waren und die Möglichkeit gegeben war, mit den Projekten in einen Dialog zur digitalen Langzeitarchivierung von Forschungsdaten zu treten. Die Interviews wurden transkribiert und den Interviewpartnern zur Korrektur und eventuellen Ergänzung vorgelegt. Die Antworten sind in dieser Studie nicht im einzelnen aufgeführt sondern werden zusammengefasst in den nachfolgenden Abschnitten dargestellt und zusammen mit den Herausforderungen diskutiert, denen sich die eScience- und Grid-Projekten stellen müssen, um den Ansprüchen digitaler Langzeitarchivierung von Forschungsdaten gerecht zu werden.

Das Feld der eScience- und Grid-Anwendung ist sehr dynamisch und entwickelt sich immer noch sehr schnell weiter. Die Expertise gibt daher den Stand der deutschen eScience- und Grid Communities im Frühjahr 2007 wieder. Spätere Entwicklungen sind, soweit redaktionell möglich, noch bis Anfang 2008 in die Expertise eingefügt worden.

Gespräche wurden mit folgenden Projekten geführt:

Grid-Projekte:

- AstroGrid-D – German Astronomy Community Grid (GACG)
- C3-Grid – Collaborative Climate Community Data and Processing Grid
- HEP-Grid – High Energy Physics Grid (Teilprojekt Theoretische Physik)
- InGrid – Innovative Grid Technology in Engineering
- MediGrid – GRID-Computing für die Medizin und Lebenswissenschaften
- TextGrid – Modulare Plattform für verteilte und kooperative wissenschaftliche Textdatenverarbeitung - ein Community-Grid für die Geisteswissenschaften

eScience-Verbundprojekte:

- eSciDoc – Plattform für Kommunikation und Publikation in wissenschaftlichen Forschungsorganisationen
- HyperImage – Bildorientierte e-Science-Netzwerke
- Im Wissensnetz – Vernetzte Informationsprozesse in Forschungsverbünden
- Ontoverse – Kooperatives vernetztes Wissensmanagement im Bereich der Life Sciences
- SYNERGIE – Verknüpfung von Informationen und Wissen durch innovative Informationstechnologie
- WIKINGER – WIKI Next Generation Enhanced Repository

[12] Siehe Neuroth et al. (2007) für eine Übersicht der im Rahmen der D-Grid-Initiative geförderten Projekte.

- WISENT – Wissensnetz Energiemeteorologie

Das D-GRID Integrationsprojekt (DGI) stellte unabhängig von unserem Fragebogen Informationen zum Thema zur Verfügung.

Erste Ergebnisse der Interviews wurden im Mai 2007 auf einem nestor-Workshop im Rahmen der German eScience Conference (GES2007) in Baden-Baden vorgestellt und mit den Stakeholdern diskutiert. Nach Abschluss der Expertise sollen die Ergebnisse und daraus abgeleiteten Handlungsempfehlungen auf einem weiteren nestor-Workshop mit den Stakeholdern reflektiert und validiert werden.

Die Ergebnisse der Befragung lassen sich in fünf Themenblöcke gliedern:

- Anforderungen an die Archivtechnologie
- Umgang mit Metadaten
- Semantische Vernetzung
- Zugang zu Daten und Rechteverwaltung
- Virtuelle Organisationen

In den nachfolgenden Abschnitten dieses Kapitels werden diese Themengebiete und die Aktivitäten der befragten Projekte dargestellt und diskutiert, um daraus in Kapitel 3 Handlungsempfehlungen abzuleiten. Um den Stand und die Strategien der Projekte zur digitalen Langzeitarchivierung zu illustrieren werden einige Aussagen aus den Fragebögen wörtlich zitiert, allerdings ohne den Namen der jeweiligen Projekte zu nennen.

2.1 Herausforderung Archivtechnologie

In der öffentlichen Diskussion wird den Archivmedien, ihrer Haltbarkeit und ihrer technologischen Zukunftsfähigkeit viel Beachtung geschenkt. In der Praxis hat sich die Erkenntnis längst durchgesetzt, dass die Medienauffrischung viel mehr ein organisatorisches, als ein technisches Problem ist. Aber auch wenn viele grundlegende technische Fragen der digitalen Langzeitarchivierung in den vergangenen Jahren bereits gelöst wurden, erwachsen mit der Verbreitung von Grid-Technologie und eScience-Projekten neue technische Herausforderungen an die digitale Langzeitarchivierung. Die Datenbestände der Grid- und eScience-Projekte stellen bereits durch ihre Größe und Komplexität neue Herausforderungen (Hey und Trefethen, 2003a; Hey und Trefethen, 2003b). So ist vorhersehbar, dass die im Laufe der Langzeitarchivierung notwendigen Medien- und Formatmigrationen sehr große Ressourcen an Rechenleistung und temporärem Speicher erfordern.

Ein wichtiger Teil der entstehenden Grid-Infrastruktur ist das Daten-Grid, das einen für den Nutzer transparenten Zugriff auf Speicherressourcen erlauben soll. So wie es heute Storage Area Networks (SAN) erlauben, separate Speicherressourcen in einer virtuelle Einheit zusammenzufassen, sollen die beteiligten Daten-Grid Komponenten einen virtuellen Datenraum aufspannen. In der Praxis sind diese Ansätze bisher auf Community-Grids begrenzt, für den Zugang von eScience-Projekten zum Daten-Grid fehlen heute noch einheitliche Schnittstellen zu den angebotenen Ressourcen. Des weiteren fehlen Standards für interoperable Kataloge und Autorisierungssysteme, mit denen Datenbestände und Dienste nachgewiesen werden und die Zugriffsrechte geregelt werden (siehe auch Abschnitt 2.4 Herausforderung Rechteverwaltung).

2.1.1 Erwartete Datenmenge

Grid-Anwendungen zeichnen sich durch sehr große Datenmengen aus. In eScience-Projekten sind die erwarteten Datenmengen im Vergleich dazu wesentlich kleiner, sind jedoch durch eine hohe semantische Komplexität gekennzeichnet. Die erste Frage befasste sich mit dem erwarteten Datenvolumen, bzw. der erwarteten Anzahl digitaler Objekte. Diese Unterscheidung wurde gemacht, da die Handhabung sehr vieler Einzelobjekte unter Umständen aufwendiger ist, als die Prozessierung der gleichen Datenmenge gebündelt in wenigen homogenen Objekten.

Wie bei der Vorbereitung der Studie erwartet, ist die erwartete Datenmenge und die Anzahl der zu verwaltenden Objekte davon abhängig, ob es sich bei dem Projekt um ein Grid- oder ein eScience-Projekt handelt[13]. Im Mittel operieren die meisten Projekte mit Datenmengen im Terabyte Bereich. Sehr große Datenmengen werden meist in Projekte mit instrumentellen Datenquellen verarbeitet (10^{15} Byte), während die Datenmengen in Projekten mit einem stark ontologischem Fokus um einige Größenordnung kleiner sind (10^{10} Byte).

2.1.2 Erwartete Dauer der Archivierung

Diese Studie beschäftigt sich mit der Langzeitarchivierung von Forschungsdaten aus eScience- und Grid-Projekten. Aber stellt sich in den Projekte selbst diese Frage? Im Gespräch wollten wir erfahren, ob die Projekte eine Langzeitarchivierung ihrer Daten anstreben, aus welchen Gründen sie dies tun und wie lange der angestrebte Archivierungszeitraum ist.

Aus den Gesprächen ergab sich, dass die angestrebte Dauer der Archivierung uneinheitlich ist. Sie reicht von wenigen Jahren bis hin zum Ziel, die Daten über Jahrzehnte – und länger – zu archivieren. Die Vorgabe der Archivierungsdauer kommt dabei stets aus der Zielgruppe des jeweiligen Projekts. In vielen Fällen werden dabei die „Empfehlungen für eine gute wissenschaftliche Praxis“ der DFG, oder ihr Equivalent in anderen Wissenschafts-organisationen genannt.

Diese Empfehlungen werden jedoch nicht im gleichen Maße als verpflichtend angesehen, wie ein in manchen Disziplinen vorhandener gesetzlicher Rahmen. Teilweise wird die Archivierungsdauer auch innerhalb der Projekte nach Datenklassen differenziert. In einigen Fällen, insbesondere bei stark ontologischem Charakter des Projekts, wird der Grad der intellektuellen Schöpfungshöhe der Daten als so hoch angesehen, dass er eine Archivierung auf unbestimmte Zeit rechtfertigt.[14]

> *„Auf Grund der kurzen Halbwertszeit der Quellen (Informatik) wird eine längere Archivierung als nicht sinnvoll angesehen. Die Entwicklung des Projekts orientiert sich eng an den Nutzerbedürfnissen. Aus diesen lässt sich bisher kein Bedarf für Langzeitarchivierung ablesen. Dennoch stellt sich die Frage nach der Persistenz der Quellen und dem langfristigen Zugang zu ihnen. Möglicherweise kann hier die Grid-Technologie in Zukunft Lösungen anbieten.“*

> *„Die Arbeitsergebnisse der Wissenschaftler gelten als Kulturgut und sind auf unbeschränkte Zeit zu archivieren.“*

[13] Größenordnungen der in den Projekten erwarteten Datenmengen: Petabyte (3), Terabyte (6), Gigabyte (3), je nach Anwendung (3)

[14] Geplante Archivierungsdauer: Unbegrenzt (2), 30 Jahre und mehr (3), 10 Jahre und mehr (5), variabel nach Projektanforderungen (1), nicht definiert (2). Mehrfachnennungen waren möglich.

„Die Dauer der Archivierung ist abhängig von den Projekten, bzw. den Produkten, die darin bearbeitet werden. Teilweise gibt es für bestimmte Produkte gesetzliche Vorgaben für die Archivierung der Dokumentation. Beispiel: bei Flugzeugen oder Kraftwerken müssen die Daten so lange archiviert werden, wie es noch Flugzeuge dieses Typs gibt, bzw. das Kraftwerk noch existiert. Bei Kraftfahrzeugen müssen die Daten für 30 Jahre archiviert werden. Diese Leistung wird vertraglich zwischen dem Hersteller und dem IT-Dienstleister geregelt."

2.1.3 Auswahlkriterien für die Langzeitarchivierung

Nicht alle Daten müssen der Nachwelt erhalten bleiben. Insbesondere bei sehr großen Datenmengen muss eine Auswahl getroffen werden. Wir wollten erfahren, welche Kriterien die Projekte für die Auswahl der Daten für eine Langzeitarchivierung zu Grunde legen.

Formale Auswahlkriterien für die Archivierung und Archivierungsdauer sind nur in den Fällen formalisiert, in denen es gesetzliche Vorgaben gibt. In den meisten Fällen sind die Auswahlkriterien differenziert nach dem angenommen Wert der Daten, also mit welchem Aufwand – wenn überhaupt – die Daten wieder zu erheben wären. Dies gilt insbesondere für Daten mit einem hohen Grad an intellektueller Schöpfungshöhe und für zeitbezogene Daten aus instrumentellen Messreihen, die nicht wiederholt werden können.[15]

„Auf Grund des hohen Grades der intellektuellen Schöpfungshöhe wird zur Zeit alles archiviert, Auswahlkriterien werden ggf. später entwickelt. Wichtig ist es, einen möglichst hohen Grad an Nachnutzbarkeit zu erreichen."

Im Gegensatz dazu wird der Wert von Zwischenprodukten aus Prozessierungsketten instrumenteller Daten oder von Simulationen als gering angesehen und somit sind diese meist nicht zur Archivierung vorgesehen.

„Es werden nur Ausgangs- und Endprodukte der Rechnungen archiviert, Zwischenprodukte der Prozessierungskette werden aus Kostengründen nur vorübergehend gespeichert. Das Problem der Migration auf neue Datenformate und neue Systemumgebungen ist noch nicht vollständig gelöst, so dass zur Zeit noch Probleme in der Nachnutzung vorhandener Daten aus inzwischen abgelösten Systemen bestehen."

2.1.4 Erwartete Datei- und Medientypen

Der Impuls, sich mit der digitalen Langzeitarchivierung von Forschungsdaten zu befassen, kam ursprünglich aus dem Kreis der Gedächtnisorganisationen und wurde im Rahmen von nestor bereits in zwei Expertisen untersucht (Severiens und Hilf, 2006a; Severiens und Hilf, 2006b). Aus dem historischen Kontext von nestor heraus lag der Schwerpunkt der Arbeit zur digitalen Langzeitarchivierung bisher auf textbasierten Formaten. Es besteht jedoch nun Forschungsbedarf zu Fragen der Formaterhaltung abseits von textbasierten Formaten. Erschwerend kommt hinzu, dass in vielen Fällen wird die Nachhaltigkeit von technischen Formaten nicht bedacht wird. Format und Informationsmodell der Daten werden oftmals nicht dokumentiert. Im Betrieb eines Projektes mag es legitime Gründe geben, auch Formate zu nutzen, deren Archivierbarkeit nicht nachhaltig gegeben ist (Lyon, 2007). Spätestens jedoch bei der Überführung der Daten in ein digitales Langzeitarchiv stellt sich jedoch die Frage

[15] Projekten, in denen formale Auswahlkriterien formuliert sind: Projektinterne Regelung (4), gesetzliche Regelung (3). Mehrfachnennungen waren möglich.

nach der Tauglichkeit der im Projekt verwendeten Formate für die digitale Langzeitarchivierung.

Projekte an der Schnittstelle zwischen eScience- und Grid-Projekten machen auf Defizite der existierenden Grid-Dienste aufmerksam. Zwischen potenziellen Anwendern von des Daten-Grid in eScience-Projekten und den Anbietern von Grid-Leistungen besteht heute noch eine mangelnde Übereinstimmung bei den Datenformaten, die im Daten-Grid abgelegt werden können. Ein objektorientiertes Datenmodell, wie es von eScience-Projekten gewünscht wird und vom OAIS-Referenzmodell vorgeschlagen wird, wäre eine mögliche Entwicklungsrichtung.

> *„Forschungsbedarf besteht bei der Adaption der Grid-Technologie von ihrem derzeit Daten-orientierten Modell zu einem Objekt-orientierten Modell, das mit Repository-Systemen wie DSpace oder Fedora kompatibel ist. Damit könnten Grids und Repositories besser miteinander genutzt werden. Derzeit müssen Repository-Funktionen im Globus Toolkit für die Verwendung im Grid nachgebaut werden."*

Die Befragung ergab, dass, sofern die Dateiformate nicht disziplinspezifisch sind, bild- und textbasierte Formate überwiegen. Auffallend ist, dass in den Projekten keine Vorgaben zu den verwendeten Dateiformaten gemacht werden, die sich an den Kriterien der Archivfähigkeit von Dateiformaten orientieren. Die technischen Aspekte der digitalen Langzeitarchivierung, wie z.B. die Auffrischung oder Migration von Medien, werden nicht als besondere Herausforderung gesehen.

Die Nutzung des Daten-Grid zur digitalen Langzeitarchivierung wird in vielen Projekten als nicht langfristig vertrauenswürdig eingeschätzt. Denkbar ist der Einsatz des Daten-Grid für viele in einer Cache-Funktion, um kurzfristig zusätzlichen Speicher nutzen zu können. Interessant erscheit für viele die Möglichkeit er einheitlichen Regelung von Zugang- und Zugriffsrechten (Single Sign-On) zu den angebotenen Ressourcen zu realisieren, allerdings wird die Rechteverwaltung im Grid als noch nicht genügend fein in ihrer Granularität angesehen. Erwartet wird auch die Auslagerung rechenintensiver Archivprozesse, z.B. im Archive-Ingest oder bei einer Formatkonversion. Aus der Sicht der Anwender sind die angebotenen Grid-Werkzeuge jedoch noch nicht stabil genug, d.h. die verwendeten Standards sind noch zu starken Veränderungen unterworfen, als dass sie im Produktionsbetrieb eingesetzt werden könnten.

> *„Die angebotenen Grid-Werkzeuge sind noch nicht stabil und können bislang nicht in einer Produktionsumgebung installiert werden. Aus diesem Grund ist auch nach anderthalb Jahren Prüfung noch keine Aussage möglich, ob der Einsatz von Grid-Technologien hier neue Lösungsansätze bietet. Die fehlende Benutzerfreundlichkeit ist eine Hemmschwelle für den Einsatz von Grid-Technologien im Alltag eines Datenzentrums."*

Besonders im Umfeld von Grid-Projekten spielen Anwendungen zur Prozessierung und Visualisierung von Daten eine wichtige Rolle. Aus diesem Grund erwächst das Bedürfnis nach einer Langzeitarchivierung von Anwendungen, bzw. des Quellcodes der Anwendungen. Eine offene Frage ist, ob zu einem Zeitpunkt in der Zukunft eine Abspielumgebung besteht, die in der Lage ist, die archivierte Anwendung, bzw. den archivierten Code auszuführen.

Derzeit konzentrieren sich die Entwicklungen in den Grid-Projekten auf Compute-Dienste, die Bereitstellung und Nutzung eines Daten-Grid steckt noch in den Anfängen und ist auf eine

gemeinsame Nutzung von Datenspeicher innerhalb eines Community-Grid begrenzt. Wenn aber, wie das Selbstverständnis von D-GRID es vorsieht, bei Spitzenlast transparent externe Ressourcen genutzt werden sollen, so ist auch eine Virtualisierung des Datenraums notwendig. Die Möglichkeit der Virtualisierung von Ressourcen würde auch die Föderation von geografisch verteilten Archiven deutlich vereinfachen (Hitchcock et al., 2007).

Wie eingangs geschildert, unterscheiden sich Grid-Projekte von eScience-Projekten in den meisten Fällen dadurch, dass der in Grid-Projekten bearbeitete Datenraum wesentlich homogener strukturiert ist, als dies in eScience-Projekten der Fall ist. Aus der Notwendigkeit heraus, heterogen strukturierte Datenobjekte in einer gemeinsamen Umgebung verwalten zu müssen, besteht bei eScience-Projekten die Nachfrage nach einer Adaption der Grid-Technologie vom bisher Daten-orientierten Modell zu einem objektorientierten Modell. Dieser Abstraktionsschritt würde es auch vereinfachen, das die Referenzarchitektur des OAIS-Archivmodells und Datenmodelle in Community-Grids zusammen zu führen.

Die komplexen Datenprodukte aus eScience-Projekten erfordern die Entwicklung eines Vorschauformats auf komplexe Datenprodukte.

> *„Forschungsbedarf besteht bei der Entwicklung eines neuen Formats für die Vorschau auf Datenprodukte. Das bisher gängige Quick-Look Format kann mehrdimensionale Objekte nicht handhaben (z.B. Monatsserien eines Produkts). Dieses Format muss letztlich standardisiert werden. Zusätzlich werden effiziente Verfahren für die Interpolation von Raum und Zeitdimensionen in Datenprodukten benötigt.“*

Dabei darf der Bedarf an Rechenressourcen für die Erstellung der Vorschau nicht unangemessenen hoch sein. Eine besondere Herausforderung sind dabei effiziente Verfahren für die Interpolation von Raum-, Zeit- oder anderen Dimensionen, die für die Auswahl von Teilmengen aus größeren Datenbeständen notwendig sind. Neben der Auswahl von Teilmengen aus großen Datenbeständen muss auch deren Referenzierbarkeit ermöglicht werden.

Wie in der einleitenden Definition des Begriffs geschildert, ist es ein Ziel von eScience die Möglichkeiten räumlich verteilten Arbeitens zu nutzen. Dazu gehört auch, dass die eScience – Arbeitsumgebung auf mehrere Datenquellen zugreifen kann. In der Praxis gestaltet sich die Erfüllung dieses Anspruchs noch schwierig, weswegen Nutzer einheitliche Schnittstellen und Protokolle für den Zugang zu Archiven und Interoperabilität zwischen Archiven wünschen.

Die Archivierung abseits von textbasierten Formaten ist nicht ein spezifisches Problem der Grid- und eScience-Projekte, wurde aber eher von eScience-Projekten genannt. Bisher war der Fokus für nicht textbasierte Formate jedoch auf Multimediaobjekten. Die entsprechenden Medienformate sind im allgemeinen gängige, industriell normierte Formate. Dass ein Format weit verbreitet ist, bedeutet jedoch nicht, dass es allen Kriterien für ein archivsicheres Datenformat genügt (Lormant et al., 2005). Der Zielkonflikt zwischen heute gebräuchlichen Formaten und archivfähigen Alternativen ist eine Herausforderung für das Management digitaler Langzeitarchivierung, die über den Zeithorizont aktiver Projekte hinaus geht und deshalb heute selten berücksichtigt wird.

> *„Forschungsbedarf besteht bei Fragen der Formaterhaltung abseits von Textbasierten Formaten. In vielen Fällen wird die Nachhaltigkeit von technischen Formaten nicht bedacht. Format und Konzept der Daten werden oftmals nicht dokumentiert. In der Zusammenarbeit mit den Wissenschaftlern sollen Community Initiativen jedoch nicht*

durch Standardisierungsprozesse blockiert werden, denn das Projekt soll Wissenschaft ermöglichen, nicht Wissenschaft verhindern.“

Insbesondere im Bereich der Community-Grids entstehen große Datenbestände in anderen, binären Formaten. Insbesondere in Grid-Projekten werden Datenprodukte verworfen, wenn die Archivierung mehr Ressourcen erfordern würde, als die Reproduktion der Datenprodukte. Auf lange Zeiträume hin tritt jedoch das Problem auf, dass die Anwendungen und Plattformen zur Verarbeitung dieser Daten nicht mehr existieren. Für die Langzeitarchivierung dieser Daten reicht daher eine reine Bit-Stream-Preservation nicht aus, denn zusätzlich müssen auch die Anwendungen archiviert werden, die für die Verarbeitung und Präsentation der Daten notwendig sind. Darüber hinaus ist es auch wahrscheinlich, dass die Hardware-Plattformen, die für die Ausführung der Anwendungen notwendig wären, zu einem Zeitpunkt in der Zukunft nicht mehr existieren, da sie als veraltet ausgemustert wurden, und nun emuliert werden müssen.

2.1.5 Langzeitarchivierung von Forschungsdaten - Best Practice Beispiele

Folgende Best Practice Beispiele wurde von Projekten genannt:

- ICSU Weltdatenzentren WDCC und WDC-MARE
- European Centre for Medium-Range Weather Forecasts
- Sloan Digital Sky Survey
- Centre de Données astronomiques de Strasbourg
- Missionsdaten der NASA, NOAA
- Bildverarbeitung in der medizinischen Versorgung
- Arts and Humanities Data Service
- Oxford Text Archive
- DANS

Die von den Projekten genannten Best Practice Beispiele sollen hier, zusammen mit einigen weiteren, kurz dargestellt werden.

ICSU Weltdatenzentren WDCC und WDC-MARE
Das World Data Center for Climate (WDCC)[16] und das World Data Center for Marine Environmental Sciences (WDC-MARE)[17] betreiben seit vielen Jahren erfolgreich die Archivierung von Forschungsdaten. Über die reine Archivierung hinaus betreiben beide Weltdatenzentren auch Datenportale, die in ihrer Nutzerfreundlichkeit als vorbildlich gelten dürfen. Beide Weltdatenzentren sind am DFG-Projekt „Publikation und Zitierbarkeit wissenschaftlicher Primärdaten“ (STD-DOI)[18] beteiligt. Über das in STD-DOI entwickelte System zur Datenpublikation versehen sie die von ihnen veröffentlichten Daten mit persistenten Identifikatoren (DOI und URN) und machen die veröffentlichten Datensätze damit dauerhaft findbar und zugänglich, eine Grundvoraussetzung für deren Zitierbarkeit (Brase, 2004; Klump et al., 2006). Zusätzlich werden ausgewählte Datensätze auch über den Katalog der Technischen Informationsbibliothek Hannover (TIBORDER) veröffentlicht. WDC-MARE stellt seinen Datenkatalog auch über maschinenlesbare Schnittstellen, z.B. OAI-PMH, für Datenportale zur Verfügung (Schindler et al., 2007).

[16] WDCC: http://wdc-climate.de/
[17] WDC-MARE: http://www.wdc-mare.org/
[18] STD-DOI: http//www.std-doi.de/

European Centre for Medium-Range Weather Forecasts (ECMWF)
Das European Centre for Medium-Range Weather Forecasts (ECMWF)[19] ist eine internationale Organisation, die gemeinsam von 28 Staaten betrieben wird. Es vertreibt seit 1979 mittelfristige Wettervorhersagen. Mit zu seinen Aufgaben gehört es, die meteorologischen Daten, auf denen die Modelle beruhen, die gerechneten Vorhersagemodelle, sowie die Vorhersagemodelle selbst zu archivieren und zugänglich zu machen. Allerdings wird kritisiert, dass der Zugang zu den Daten benutzerfreundlicher gestaltet sein könnte. Daten werden auch über Webservices übertragen, die jedoch nicht XML-Codiert sind.

Sloan Digital Sky Survey (SDSS)
Die Himmelskartierung durch den Sloan Digital Sky Survey (SDSS)[20] ist die ehrgeizigste Himmelskartierung, die bisher unternommen wurde. In der ersten Phase des Projekts von 2000 bis 2005 wurden ein Viertel des Himmels kartiert und ein dreidimensionales Modell des Kosmos mit etwa einer Million Galaxien und Quasaren erstellt. Die Daten werden in jährlichen Veröffentlichungen über den freigegeben und sind im Internet über den SDSS SkyServer frei zugänglich. Die Daten sind nach unterschiedlichen Produktgruppen geordnet. Für den Zugang zu den Daten stehen unterschiedliche Kataloge zur Verfügung.

Centre de Données astronomiques de Strasbourg (CDS)
Das Centre de Données astronomiques de Strasbourg (CDS)[21] gehört zu den ältesten digitalen Datenarchiven und besteht seit Mitte der 1970er Jahre. Einen deutlichen Aufschwung erfuhr das CDS durch das Internet, da es durch neue Dienste immer weitere Nutzerkreise erschließen konnte (Genova et al., 2005). Bemerkenswert ist der Katalog des CDS der durch die Simbad Referenzdatenbank gleichzeitig auch die synonymen Bezeichnungen astronomischer Objekte verwaltet. Zusätzlich werden in Simbad auch Basisdaten zu den katalogisierten Objekten gespeichert und die dazu gehörige Literatur referenziert. CDS bietet Schnittstellen zu seinen Datenbeständen an, darunter seit 2002 auch XML-Webservices auf der Basis des SOAP-Protokolls.

Missionsdaten der NASA und NOAA
Als Teil ihrer Aufgaben betreiben die National Aeronautics and Space Administration (NASA)[22] und die National Oceanic and Atmospheric Administration (NOAA)[23] seit einigen Jahrzehnten Fernaufklärungsmissionen und Sensornetzwerke, aus denen enorme Mengen an Daten und Datenprodukte hervorgehen. Beide Organisationen bearbeiten sehr weite Aufgabenfelder. Daten aus Erdbeobachtungsmissionen sind zu weiten Teilen über Webportale zugänglich und können zum Teil auch über Webservices direkt abgerufen werden. Während NOAA keinen zentralen Katalog betreibt, sondern sie e nur über den Global Change Master Directory (GCMD) und ihre Fachportale veröffentlich, katalogisiert die NASA ihre Datenbestände und macht sie über GCMD und Federation Interactive Network for Discovery (FIND) verfügbar. Mit zu FIND gehört auch der EOS Data Gateway (EDG). Beide Organisationen beschäftigen sich intensiv mit der digitalen Langzeitarchivierung von Daten aus der Erdbeobachtung und wie diese für eine Nachnutzung verfügbar gemacht werden können.

[19] ECWMF: http://www.ecmwf.int/
[20] SDSS: http://www.sdss.org/
[21] CDS: http://cdsweb.u-strasbg.fr/
[22] NASA:
[23] NOAA: http://www.noaa.gov/

Arts and Humanities Data Service (AHDS)
Der Arts and Humanities Data Service (AHDS)[24] wurde 1996 eingerichtet, um elektronische Ressourcen aus Forschung und Lehre in den Geistes- und Sozialwissenschaften zu sammeln. Der Datenbestand wird über einen Online-Katalog veröffentlicht. Der AHDS-Katalog verweist direkt auf die im Katalog nachgewiesenen Quellen. Diese können Sammlungen von historischen Artefakten sein, aber auch Datensätze aus der historischen und sozialwissenschaftlichen Forschung. Ob die Daten auch über Webservices erreichbar sind, hängt von den einzelnen Sammlungen ab, der AHDS unterstützt den Einsatz von Webservices in allen Prozessen der digitalen Langzeitarchivierung. Der AHDS und seine Datenzentren beteiligen sich an einer großen Anzahl von Projekten, in denen offene Fragen der digitalen Langzeitarchivierung bearbeitet werden.

Oxford Text Archive
Das Oxford Text Archive (OTA)[25] ist eines der am AHDS beteiligten Zentren. Im AHDS ist das OTA zuständig für die Archivierung von hochwertigen elektronischen Dokumenten zu Literatur, Sprachen und Linguistik für Forschung und Lehre. Neben der Archivierung von Texten hat das OTA auch das Ziel, die Dokumentation von elektronischen Dokumenten in der geisteswissenschaftlichen Forschung zu standardisieren und zu verbessern, um deren Qualität und Nachnutzbarkeit zu erhöhen. Die Standardisierung der Textformate und –beschreibungen steht auch im Dienste der digitalen Langzeitarchivierung. Die Archivierung von Daten wird aktiv unterstützt und entsprechende Dienste potenziellen Nutzern angeboten.

Data Archiving and Networked Services (DANS)
Data Archiving and Networked Services (DANS[26]) ist eine Einrichtung der Königlich Niederländischen Akademie der Wissenschaften und damit beauftragt, Forschungsdaten aus den Geistes- und Sozialwissenschaften zu archivieren und verfügbar zu machen. DANS arbeitet eng mit den niederländischen Forschungsinstituten und internationalen Datenanbietern zusammen. Dabei sind die Daten nicht unbedingt zentral bei DANS gespeichert, sondern können auch in institutionellen Repositorien liegen, denn neben dem Auftrag, Forschungsdaten zu archivieren, hat DANS auch den Auftrag, Qualität, Zugang und Nutzbarkeit externer Datenquellen zu zertifizieren. Die DANS Richtlinien orientieren sich am OAIS-Referenzmodell und u.a. auch am nestor „Kriterienkatalog für vertrauenswürdige digitale Langzeitarchive“ (Dobratz et al., 2006). Im Rahmen von DANS werden auch durch Forschungsprojekte Fragen der Zukunft digitaler Langzeitarchivierung untersuch, wie z.B. der Einsatz von Grid-Techologie oder persistenter Identifikatoren. Ein Teil der Archivprozesse in DANS sind als Webservice verfügbar.

UK Data Archive (UKDA)
Das UK Data Archive (UKDA)[27] hat die Aufgabe, Forschung und Lehre in Sozial- und Geisteswissenschaften durch die Erfassung von Daten und Datenmanagement zu unterstützen, die Ressourcen und Dienste weiter zu entwickeln und der wissenschaftlichen Gemeinschaft und der Öffentlichkeit bekannt zu machen. Das UKDA hat den Auftrag, die Daten und ihre Dokumentation auf lange Zeit zu erhalten und zugänglich zu machen. Mit zum Auftrag des UKDA gehört auch die Beobachtung und Bewertung der technischen Entwicklung und ihrer Auswirkung auf das Management der Datenerhaltung und –migration.

[24] AHDS : http://ahds.ac.uk/
[25] OTA : http://ota.ahds.ac.uk/
[26] DANS: http://www.dans.knaw.nl/en/
[27] UKDA: http://www.data-archive.ac.uk/

Der Data and Support Service des UKDA hat die Aufgabe, Daten für die digitale Langzeitarchivierung zu identifizieren, in das Archiv aufzunehmen und mit Metadaten anzureichern. Des Weiteren unterstützt diese Abteilung die Nutzer beim Zugang zu den Daten und ihrer Nachnutzung.

Bildverarbeitung in der medizinischen Versorgung
Auf Grund der gesetzlichen Vorgaben und der großen Menge an Bilddaten aus der medizinischen Versorgung wird in diesem Bereich seit langem die digitale Langzeitarchivierung weiter entwickelt. Der Gesetzgeber schreibt vor, dass Bilder und Befunde, die der medizinischen Versorgung dienten, für eine Dauer von 30 Jahren archiviert werden müssen. Daten aus frühen Phasen der digitalen Bildverarbeitung haben also bereits mehrere Format- und Medienwechsel überstehen müssen. Der relativ homogene Datenraum, der im Wesentlichen aus Rasterbildern und Texten besteht, kommt der Aufgabe der digitalen Langzeitarchivierung entgegen. Die Datenmenge entspricht einer Größenordnung, wie sie auch in Grid-Projekten erwartet wird. Das Beispiel zeigt jedoch auch, wie durch gesetzliche Vorgaben ein Rahmen geschaffen wurde, in welchem die Belange der digitalen Langzeitarchivierung berücksichtigt werden müssen.

Protein Data Bank (PDB)
Die Protein Data Bank (PDB)[28] ist eine Datenbank für 3D-Strukturdaten von Proteinen, Nukleinsäuren und großer makromolekularer Komplexe. Die PDB fungiert als Repository für die Koordinaten und damit verbundenen Information von über 38000 molekularer Strukturen, die mit den Methoden der Röntgendiffraktometrie, Kernspinresonanz und Elektronenmikroskopie entschlüsselt wurden. Ziel der PDB ist es, eine zentrale Datenbasis aller veröffentlichten makromolekularen Strukturdaten aufzubauen und zu betreiben und diese Daten öffentlich und kostenlos zugänglich zu machen und den Nutzern zusätzliche Mehrwertdienste zur Verfügung zu stellen. Die PDB ist ein internationaler Verbund mehrerer biowissenschaftlicher Forschungszentren. Eine ausführliche Beschreibung der PDB und ihrer Ziele findet sich in (Berman et al., 2007).

2.2 Herausforderung Metadaten

Reine Datendateien sind ohne Beschreibung ihrer Struktur, ihrer Herkunft oder ihrer Benutzung schon nach kurzer Zeit nicht mehr nutzbar. Disziplinspezifische Beschreibungen der Daten helfen, diese zu lokalisieren, zu interpretieren und nachzunutzen. Aus diesem Grund messen wir Metadaten eine hohe Bedeutung bei. Um den Handlungsbedarf identifizieren zu können, befragten wir die Grid- und eScience-Projekte zu ihrem Umgang mit Metadaten, den verwendeten Metadaten-Standards, zur Dokumentation der verwendeten Dateitypen und zur Dokumentation der Produktion der Daten. Der "Kriterienkatalog vertrauenswürdige Langzeitarchive" (Dobratz et al., 2006) definiert Metadaten im Sinne einer digitalen Langzeitarchivierung wie folgt:

> *„Zu den Daten, die die Inhaltsinformation repräsentieren (Inhaltsdaten), können weitere Daten hinzukommen, die z.B. der Identifizierung, Auffindbarkeit, der Rekonstruktion und Interpretation oder dem Nachweis der Integrität und Authentizität sowie der Kontrolle der Nutzungsrechte dienen (Metadaten). Metadaten können zu unterschiedlichen Zeiten im Lebenszyklus digitaler Objekte entstehen (z.B. bei der Produktion, bei der Archivierung, bei der Bereitstellung für die Nutzung). Sie werden als Teile der logischen Einheit* ‚digitales Objekt' *aufgefasst und können sowohl getrennt als auch gemeinsam mit den Inhaltsdaten verwaltetet werden."*

[28] PDB: http://www.pdb.org/pdb/Welcome.do

Das OAIS-Referenzmodell dient als Vorbild bei der Bewertung der Frage, ob der Umgang mit Metadaten in den Projekten den Anforderungen der digitalen Langzeitarchivierung gerecht wird. Metadaten erfüllen hier wichtige Funktionen, sie beschreiben Datensätze mit Attributen, die von der Nutzercommunity als wesentliche Beschreibungsmerkmale angesehen werden (Descriptive Metadata), und sie dienen der Verbreitung und Auffindbarkeit von Daten in Katalogen und Katalogdiensten (Discovery Metadata). Für die Nachnutzung von Daten ist zudem wichtig, dass die verwendeten Dateiformate (Representation Metadata), die Herkunft (Provenance Metadata) und die Prozessierungsschritte (Processing Metadata) zur Erzeugung der Daten dokumentiert sind.

Von den befragten Projekten wird die Grid-Technologie als Motor für die Verbreitung einheitlicher Schnittstellen gesehen, einer Grundvoraussetzung für modulare, Service Orientierte Architekturen. Einige Projekte erwarten vom Daten-Grid als Service neue Lösungsansätze für die digitale Langzeitarchivierung.

> *„Daten werden in Zukunft immer häufiger über mehrere Institutionen verteilt sein. Standards für Schnittstellen und Metadaten werden von den Entwicklungen der Grid-Community profitieren."*

Dennoch sind Metadaten ein Thema, an dem sich die Geister scheiden, denn die Erzeugung von Metadaten wird im allgemeinen als sehr lästig wahrgenommen, währen die angebotenen Metadatenprofile als entweder über- oder unterkomplex eingeschätzt werden. Zudem fehlen in vielen Fällen noch geeignete Werkzeuge um Metadaten zu erzeugen oder zu editieren.

Überzeugende Verfahren für den Umgang mit Metadaten sind heute in Communities zu finden, die in ihren Arbeitsabläufen eine weitgehend automatisierte Erzeugung von Metadaten integriert hat. Vorbildlich im Umgang mit Metadaten sind Bereiche Klimaforschung (Kindermann et al., 2006) und Biodiversitätsforschung (Cotter et al., 2004; Fornwall, 2004). Beide Bereiche haben in den vergangenen Jahren große Fortschritte im Umgang mit Metadaten gemacht, indem sie analysierten, welche Schnittstellen zwischen Informationssystemen und Anwendungen bedient werden sollten. Die benötigten Metadaten werden weitgehend automatisch erstellt. Diese Ansätze gilt es auf andere Disziplinen zu übertragen.

2.2.1 Metadaten-Standards

Der Bedeutung von Metadaten sind sich alle befragten Projekte bewusst. Teilweise existieren jedoch noch keine in der Community akzeptierten Metadatenstandards. Die Ursache für fehlende Metadatenstandards kann darin liegen, dass ein „anonymer" und interdisziplinärer Austausch von Daten im genannten Feld eine neue Entwicklung in der wissenschaftlichen Zusammenarbeit ist. Es sind jedoch auch Fälle bekannt, in denen Metadatenschemata für eine Domäne existieren, jedoch in der Fachcommunity keine Akzeptanz finden, weil die Schemata entweder als zu einfach oder – was häufiger vorkommt – als übermäßig komplex angesehen werden. Fehlende oder inkonsistente Metadatenschemata sind jedoch ein ernsthaftes Hindernis für die digitale Langzeitarchivierung von Forschungsdaten, genauso wie fehlende Metadaten.

> *„Der [vorgeschlagene] Standard wird von der Community als zu komplex abgelehnt."*

Soweit vorhanden, folgen die in den Projekten verwendeten Metadatenprofile anerkannten Standards. Dabei handelt es sich überwiegend um Profile für Discovery Metadaten. Die

weiteste Verbreitung finden dabei einfach strukturierte Schemata, wie z.B. Dublin Core, während komplexe Schemata von den Communities oft nur zögernd akzeptiert werden, weil sie als zu kompliziert und nicht handhabbar gelten. Im allgemeinen ist ein Bewusstsein für die Bedeutung von Metadaten, ihrer Einsatzfelder und ihrer Qualität vorhanden. Die Befragung der eScience- und Grid-Projekte hat jedoch gezeigt, dass es in Bezug auf Metadaten zwei mit einander verwandte Probleme gibt: Metadatenschemata und die Erzeugung von Metadaten.

> *„Soweit Standards vorhanden sind, werden diese verwendet. Allerdings decken [die anerkannten] Metadatenprofile nicht alle [gewünschten] Metadaten-Attribute ab."*

> *„Das Metadatenformat orientiert sich an Dublin Core und hat den Anspruch möglichst offen und flexibel zu sein. Cross-walks zu anderen Metadatenstandards werden angestrebt."*

Bemerkenswert ist die Konzentration auf Katalog-Metadaten, die Beschreibung der Inhalte und die Verwaltung der Zugriffsrechte. Eine Standardisierung der Dokumentation der Herkunft von Daten, ihrer Lizenzierung, des Dateiformats oder von semantischen Verweisen wird nicht angesprochen.

> *„Forschungsbedarf besteht auf technischer Ebene bei der Interoperabilität zwischen Datenspeichersystemen. Außerdem fehlen Standards für interoperable Kataloge und Autorisierungssysteme. Fehlende Standards behindern den Austausch von Daten zwischen Grids."*

Zwar dienen Metadaten der Beschreibung von Daten und Diensten, vielfach besteht jedoch das Missverständnis, dass Metadaten „menschenlesbar" sein sollten, und dass man sich auf genau einen einzigen Standard einigen muss (Severiens und Hilf, 2006a). Dass es auch ganz anders geht, zeigen die Best-Practice Beispiele für den Umgang mit Metadaten in Abschnitt 2.2.4. Hier werden die Metadaten intern in einem eigenen Schema vorgehalten und dann je nach Anwendung in das angefragte Schema umgeformt und über das entsprechende Protokoll ausgegeben. Hier steht die Kommunikation von Maschine zu Maschine im Vordergrund, wobei die Darstellung der Metadaten für menschliche Leser auch möglich ist.

2.2.2 Metadaten zum Dateityp

Ohne Dokumentation der Dateiformate besteht die Gefahr, dass Dateien von technisch obsoleten Plattformen trotz einer erfolgreicher Migration auf eine neue Plattform nicht mehr benutzt werden können. Eine weitere Voraussetzung für die Nachnutzung von Daten ist, dass deren Herkunft und Prozessierung dokumentiert sind, denn in den meisten Fällen handelt es sich bei archivierten Daten nicht um Rohdaten. Auch im Fall, dass tatsächlich Rohdaten archiviert wurden, müssen Metadaten mit archiviert werden in denen die Herkunft der Daten festgehalten sind, z.B. instrumentelle Parameter oder Dokumentation der Zusammenstellung eines soziologischen Pannels.

Nur in wenigen Fällen wird mit Metadaten dokumentiert, wie die Daten gewonnen und bearbeitet wurden. Bei der Dokumentation der Datenherkunft und –bearbeitung haben die ontologisch orientierten eSciencen-Projekte in der heutigen Praxis meist einen Vorsprung gegenüber den Datenorientierten Grid-Projekten, da dieser Aspekt in vielen Metadatenprofilen der Grid-Projekte fehlt oder nur schwach entwickelt ist. Diese Information ist jedoch nicht zu vernachlässigen, da sie für die Nachnutzung von Daten unbedingt notwendig ist. Problematisch ist zudem, dass die genutzten Dateiformate meist nur implizit beschrieben sind, z.B. über ihren MIME-Type.

„Die Benutzung der Datei-Typen geht aus deren MIME-Type hervor."

„Die Daten selbst sind einfach da. Was damit gemacht wird, ist Sache der Dienste, die im System installiert sind. Die Eignung bestimmter Daten zu bestimmten Verarbeitungen leitet sich aus den MIME-Typen her."

In der Praxis hat sich gezeigt, dass es für eine langfristige Lesbarkeit der archivierten Dateiformate nicht ausreicht, allein deren MIME-Type zu kennen. Es ist daher notwendig, die genaue Dateistruktur zu kennen, was im Falle proprietärer, nicht standardisierter Formate ein Problem sein kann, auch wenn diese Formate eine weite Verbreitung haben. Ein Zusätzliches Problem kann die fehlende Kompatibilität zu älteren Versionen des Dateiformats sein. Alle Leistungsfähigkeit eines Langzeitarchivs mit Grid-Unterstützung ist vergebens, wenn nicht bekannt ist, wie ein vorliegendes Format in ein neues überführt wird. Aus diesem Grund ist es notwendig, dass sich die Archivbetreiber und die Projekte mit der Archivfähigkeit von Dateiformaten auseinandersetzen (Curtis et al., 2007; Lormant et al., 2005).

Wenn jedoch bekannt ist, wie ein Dateityp zu lesen und zu benutzen ist, können durch Emulationen auch Datenbestände von obsoleten Plattformen wieder genutzt werden. Dies wurde von einem der Grid-Projekte genannt.

„Metadaten zum Dateityp und zur Prozessierung gespeichert. Software-Emulationen werden genutzt, um Legacy-Daten zugänglich zu machen."

2.2.3 Metadaten zur Datenherkunft und Prozessierung

Ein verwandter Aspekt ist die Kodifizierung der Prozessierung und Herkunft der Daten, denn ohne diese Information ist eine Nachnutzung wissenschaftlicher Primärdaten kaum möglich. Bislang unterstützen nur wenige Systeme die Erfassung der Datenherkunft als Teil des wissenschaftlichen Arbeitsprozesses, bzw. werden diese Informationen nicht archiviert oder in die Metadaten integriert.

„Über den Einsatz von PREMIS[29] wird nachgedacht um die Herkunft und Prozessierung von Daten aufzuzeichnen. Die derzeitigen Prozessketten können noch nicht 1:1 abgebildet werden."

„Metadaten zu Herkunft und Prozessierung der Daten werden teilweise gespeichert. Dies ist insbesondere dort notwendig, wo entsprechende rechtliche Anforderungen bestehen."

„Die Dokumentation der automatischen und manuellen Prozessierung ist möglich, ist aber optional."

Am weitesten fortgeschritten ist die Dokumentation der Datenherkunft bei den stark ontologisch ausgerichteten eScience-Projekten.

„Die Datenherkunft wird als Prozessontologie gespeichert, ein abstraktes Modell für den Wissensschöpfungsprozess und dessen Aktivitäten. Daten werden zum Teil aus bestehenden Datenbeständen extrahiert, bzw. kooperativ von den Forschern erstellt.

[29] Siehe Abschnitt 1.1

Die Prozessierung der Daten erfolgt in einer gemeinsamen Infrastruktur über webbasierte Anwendungen und Office-Software."

„Ins System importierte Daten werden mit Identifikatoren versehen und ändern sich nach Import nicht mehr, d.h. die Daten werden versioniert abgelegt. Jede Version bleibt erhalten, ist als Version eines Dokuments erkenntlich und kann gesondert adressiert werden. Prozesse, die Metadaten erzeugen, legen diese in eigenen Objekten ab, die einen Bezug zu den Ursprungsdaten aufweisen und ebenso versioniert werden."

Allgemein bieten Grid-Projekte, in denen Prozesse über Workflows gesteuert werden, gute Voraussetzungen für die strukturierte Archivierung von Metadaten zur Herkunft und Prozessierung von Daten, da hier beides explizit kodiert wird. Allerdings besteht hier noch Forschungsbedarf, da Workflow-Beschreibungen und Workflow-Engines sich noch sehr an den Bedürfnissen des eBusiness orientieren (Barga und Gannon, 2007).

2.2.4 Metadaten - Best Practice Beispiele

In der Befragung der Projekte wurde kein Best Practice Beispiel für den Umgang mit Metadaten genannt, obwohl es z.B. in der Klimaforschung und Biodiversitätsforschung durchaus vorbildliche Ansätze gibt. Speziell als Lösung für die Dokumentation von Dateiformaten, die über den MIME-Type hinaus gehen, ist die Arbeitsgruppe „Data Format Description Language" des Open Grid Forum interessant.

PANGAEA Publishing Network for Geoscientific & Environmental Data

Das PANGAEA Publishing Network for Geoscientific & Environmental Data[30] ist ein Informationssystem für die Erd- und Umweltwissenschaften, das gemeinsam vom Alfred-Wegener-Institut für Polar- und Meeresforschung und dem Zentrum für Marine Umweltwissenschaften (MARUM) der Universität Bremen betrieben wird. PANGAEA ist auch die technische Plattform des World Data Center for Marine Environmental Sciences (WDC-MARE). In PANGAEA werden Metadaten nach einem internen Schema vorgehalten und dann nach Bedarf in verschiedene gängige Metadatenstandards transformiert und können über eine Reihe von Protokollen an Client-Systeme oder externe Anwendungen ausgegeben werden (Huber und Schindler, 2007).

Data Format Description Language (DFDL)

Data Format Description Language (DFDL)[31] ist eine Arbeitsgruppe des Open Grid Forum (OGF)[32] mit dem Ziel, eine XML-basierte Beschreibungssprache für strukturierte binäre oder Zeichenbasierte Dateien und Datenströme zu einwickeln, damit deren Format, Struktur und Metadaten offen gelegt werden können. Die Möglichkeit, Datenformate mit mehr als ihrem MIME-Type standardisiert zu beschreiben, ist für die digitale Langzeitarchivierung von wissenschaftlichen Primärdaten unbedingt notwendig. Eine Offenlegung und Dokumentation von Dateiformaten würde auch die Bewertung ihrer Eignung zur Langzeitarchivierung und möglicher Formatmigrationen unterstützen.

2.3 Herausforderung Semantic Web

Bei der Konzeption der Studie wurde davon ausgegangen, dass es eine Dualität von Grid-Projekten mit großen Datenvolumina und eScience-Projekten mit hoher semantischer Komplexität gibt. Generell zeichnen sich Grid-Projekte durch eine niedrigere semantische Komplexität aus, während die semantische Komplexität von eScience-Projekten immer sehr

[30] PANGAEA: http://www.pangaea.de
[31] DFDL: http://forge.gridforum.org/sf/projects/dfdl-wg
[32] OGF: http://www.ogf.org/

hoch ist. Bei einigen Grid-Projekten handelt es sich jedoch Mischformen, die gleichzeitig relativ hohe Datenvolumina und einen hohen Grad an Vernetzung zwischen den Objekten handhaben müssen.

Darüber hinaus interessierte uns, wie das weitergehende Potenzial des Semantic Web in den Projekten genutzt wird insbesondere bei der Vernetzung mit physischen Objekten („Internet der Dinge“) und für die Erfassung von implizitem Wissen und Prozesswissen.

Ein Teil der eScience-Projekte sind per Definition stark ontologisch orientiert und somit ist die Kodifizierung von Semantik und Ontologien ihre *raison d'être*. Bei weniger ontologisch orientierten eScience-Projekten und bei Grid-Projekten steht jedoch eher die Datenintegration im Vordergrund, die auch eine semantische Integration bedeutet. Für eine automatische Datenintegration ist es jedoch notwendig, Ontologien zu entwickeln, mit denen sich die semantischen Modelle der unterschiedlichen Datenquellen auf ein gemeinsames semantisches Modell abbilden lassen.

Die semantische Beschreibung in menschenlesbarer, als auch in maschinenlesbarer Form soll es einfacher machen, Dienste und Datenquellen zu finden und in eScience- oder Grid-Anwendungen zu integrieren. Diese Erweiterung des bestehenden Grids um eine semantisch konsistente Beschreibung wird auch als Semantic Grid bezeichnet (De Roure et al., 2005). Dabei handelt es sich jedoch nicht um Wissensmanagement in der „Wissensschicht“ der eScience-Anwendungen sondern um ein Integrationswerkzeug in der Middleware-Schicht der Community-Grids. Eine semantisch konsistente Beschreibung der Grid-Ressourcen, sowohl Daten als auch Dienste, wäre ein wertvoller Beitrag zu einer nachnutzbaren Langzeitarchivierung von Daten aus eScience- und Grid-Projekten.

> *„Wir sehen Forschungsbedarf bei der Zusammenführung von Grid- und Semantic-Web-Technologien.“*

Aus Sicht der eScience-Projekte ist die semantische Integration von Grid-Diensten noch ungenügend. Dieser Umstand behindert eine die Integration verteilter, heterogener und dynamischer Quellen und Dienste. Es besteht bedarf an einer Integration von Grid- und Semantic Web-Technologien, wie sie das Semantic Grid vorsieht.

2.3.1 Semantische Beziehungen zwischen Daten

Insbesondere für die Nutzung von Daten mit Werkzeigen des eScience ist es interessant, wenn auch die Beziehungen zwischen Datensätzen aufgezeichnet und archiviert werden. Nach Aussage der Gesprächspartner sind Beziehungen zwischen gespeicherten Objekten nicht in allen Projekten relevant. In den Fällen, in denen Beziehungen zwischen Objekten hergestellt werden sollen, werden zwei sehr unterschiedliche Wege gewählt. Im einen Fall werden Beziehungen durch Objekt-Identifier impliziert. Im anderen Fall werden die Beziehungen explizit formalisiert und in RDF oder OWL kodiert. In Einzelfällen kommen auch andere Domänenspezifische Ontologie-Schemata zum Einsatz.[33]

Im Projekt eSciDoc wird in der Erfassung semantischer Beziehungen zwischen Daten ein hoher Stellenwert beigemessen:

[33] Projekte, die semantische Beziehungen zwischen Datensätzen speichern: explizite Verweise (8), implizite Verweise (6). Mehrfachnennungen waren möglich.

„Semantische Beziehungen zwischen Daten werden in RDF repräsentiert, die Entwicklung einer eSciDoc-Ontologie wird diskutiert. Das Ziel ist auf jeden Fall ein Management der Heterogenität, eine allgemein gültige Lösung wird es kaum geben. Nötig wäre in diesem Fall ein Mapping der Ontologien aufeinander um neue semantische Sichten auf die Daten zu ermöglichen."

Die Entwicklung und Pflege solcher Ontologien ist sehr aufwändig. Einen interessanten Ansatz verfolgt hier das Projekt ONTOVERSE:

„Semantische Beziehungen zwischen Daten werden in RDF oder OWL repräsentiert. Zusätzlich können die Nutzer frei Tags vergeben. Die Tags werden auf die Ontologie abgebildet und ggf. in die Ontologie integriert."

Das Vorgehen im Projekt ONTOVERSE ist ein sehr interessanter Ansatz, die in vielen Bereichen noch vorherrschende Kluft zwischen standardorientierter und nutzerorientierter Entwicklung zu überwinden.

2.3.2 Semantische Vernetzung der Daten mit anderen Objekten

Forscher wollen vorhandene Daten nachnutzen. Um Inhalt, Qualität und damit die Nachnutzbarkeit der Daten einschätzen zu können, orientieren sie sich an der Interpretation der Daten in der wissenschaftlichen Fachliteratur. Zudem gibt es Fälle, z.B. in der Biologie, Geologie, oder Archäologie, in denen es für die Forscher interessant ist, das Objekt, an dem die Daten erhoben wurden, identifizieren zu können. Uns interessierte daher zu erfahren, ob in den Projekten semantische Verbindungen zwischen Veröffentlichungen, Daten und Forschungsmaterialien mit verwaltet werden (Semantic Web und Internet der Dinge).

Beziehungen zwischen Literatur und Daten werden in einigen Projekten verwaltet. In anderen Projekten ist diese Art von semantischer Vernetzung nicht relevant. Die Bandbreite ist jedoch recht groß und reicht von der Aussage, dass dies zum Kern des Projektes gehöre, über die Aussage, dass diese Art von Vernetzung als zu aufwendig angesehen wird, bis hin zu Fällen, in denen sie im Projekt nicht anwendbar ist. Die Entscheidung über die Erfassung von Beziehungen zwischen Daten und Literatur ist somit eng mit den Zielen des Projekts verbunden.

In wenigen Fällen werden bisher auch Beziehungen zwischen Daten und physischen Objekten erfasst.

„Die semantische Vernetzung von Literatur und Daten wird im Projekt verwaltet. Auch externe Quellen können eingebunden werden, falls notwendig über Proxyobjekte, die mit den notwendigen Metadaten versehen sind."

„Semantische Beziehungen zwischen Daten und Veröffentlichungen werden [im Projekt] verwaltet. Zusätzlich wird angestrebt die Beziehung zu Biomaterialien nach dem Vorbild des Shared Pathology Information Network (SPIN)[34] mit zu erfassen."

Ein vergleichbares Ziel verfolgt das Projekt System for Earth Sample Registration (SESAR)[35]. Hier werden geologische Proben mit einer International Geo-Sample Number (IGSN) versehen, um Analysen und deren Interpretationen in der Literatur eindeutig den

[34] SPIN: http://spin.nci.nih.gov/
[35] SESAR: http://www.geosamples.org/

Proben zuordnen zu können, an denen die Daten gemessen wurden (siehe auch Abschnitt 2.3.4 Semantic Web - Best Practice Beispiele).

2.3.3 Umgang mit implizitem Prozesswissen

In die Erstellung der Daten fließt in vielen Fällen einiges an implizitem Prozesswissen ein, das über Herkunft und Prozessierung der Daten hinaus geht. Nicht alle Gesprächsteilnehmer waren sich der Bedeutung implizitem Prozesswissens bewusst. In einigen Projekten wird Prozesswissen jedoch als relevant angesehen und deshalb auch archiviert. Dies gilt insbesondere für Projekte mit stark ontologischer Ausrichtung. Allerdings stößt die Dokumentation von implizitem Prozesswissen auch auf Vorbehalte, wenn in diesem Wissen der Wettbewerbsvorteil des Akteurs liegt.

> *„Die Dokumentation impliziten Wissens hat bei großen Unternehmen einen hohen Stellenwert. Klein- und mittelständischen Unternehmen hätten Bedarf, aber scheuen sich, ihren Wettbewerbsvorteil, den sie durch implizites Wissen haben können, zu dokumentieren um Industriespionage zu erschweren. Produktionsprozesse lassen sich noch nicht kopieren."*

> *„In einem Teilprojekt des Projekts geht es genau darum, implizites und Prozesswissen zu dokumentieren. Das Teilprojekts ist jedoch noch nicht begonnen worden, da noch hohe rechtliche Hürden zu überwinden sind, um den Anforderungen des Patent- und des Haftungsrechts gerecht zu werden."*

Am weitesten ist die Dokumentation impliziten Wissens in den Projekten ONTOVERSE und SYNERGIE fortgeschritten. In ONTOVERSE wird Prozesswissen in einem Ontology Requirements Specification Document (ORSD) aufgezeichnet. Das ORSD ist Wiki-basiert und dokumentiert Ziel, Umfang, Kompetenz, Erfolge und Fehler der eingesetzten Ontologie. In SYNERGIE wird mit den Ergebnissen stets ausgewiesen, welches Verfahren angewendet wurde. Auch die Arbeit der Nutzer wird mit protokolliert, um die Arbeitsweisen der Nutzer kennen zu lernen und Routineoperationen ggf. als vorprozessiertes Produkt anzubieten.

2.3.4 Semantic Web - Best Practice Beispiele

Shared Pathology Information Network (SPIN)
Das Ziel des Shared Pathology Information Network ist es, Forschern internetbasierte Werkzeuge an die Hand zu geben, um für ihre Forschung geeignete menschliche Gewebeproben zu finden. Das Netzwerk stellt anonymisierte Informationen über die Verfügbarkeit von Pathologieproben bereit, die für die vom Forscher angegebene Fragestellung geeignet sein könnten.

System for Earth Sample Registration
Das System for Earth Sample Registration ist ein zentraler Dienst für die Vergabe und Verwaltung von Identifikatoren (International Geosample Number, IGSN) für geowissenschaftliches Probenmaterial. Die Verwendung von IGSNs soll die Bezeichnung von Proben systematisieren um uneindeutige Probennamen, wie sie heute häufig in der Literatur vorkommen, zu vermeiden. Mit der Vergabe von IGSNs baut SESAR einen globalen Katalog von geowissenschaftlichem Proben auf. Das Integrated Ocean Drilling Program plant, analytische Daten mit den IGSNs der Proben, an denen diese gemessen wurden, zu verknüpfen.

2.4 Herausforderungen Zugang zu Daten und Rechteverwaltung

In der Diskussion über digitale Bibliotheken und den offenen Zugang zu wissenschaftlichem Wissen ist auch der Zugang zu Daten und deren Austausch unter Wissenschaftlern Gegenstand des Diskurses über die Zukunft der Forschung geworden. Aus diesem Grund haben wir gefragt, ob die Daten auch für Dritte zugänglich gemacht werden (Data sharing).

Der Zugang zu Daten wird in allen Projekten begrüßt, aber nicht immer auch umgesetzt[36]. Ob er auch in den Projekten selbst umgesetzt wird, hängt auch davon ab, ob es Gründe gibt, die einem offenen Zugang zu Daten entgegen stehen. In erster Linie handelt es sich um rechtliche Beschränkungen, die dem Schutz von Daten von Personen oder Unternehmensdaten dienen. Aber auch andere Gründe, die aus dem Schutz der Persönlichkeitsrechte oder Urheberrechte motiviert sind, können eine Rolle spielen.

> *„Grundsätzlich wird der offene Zugang zu Daten unterstützt. Die Praxis muss sich jedoch aus der Community heraus entwickeln. Die beteiligten Wissenschaftler sollen durch einen möglichst großen Kooperationsgewinn vom Nutzen des offenen Zugangs zu Daten überzeugt werden."*

> *„Wichtigstes Ziel des Projekts ist der Austausch von Daten zwischen Forschergruppen. Überwiegend handelt es sich dabei um bilateralen Austausch von Daten zwischen Forschergruppen. Der Offene Zugang zu Daten wird diskutiert und allgemein auch akzeptiert. Stellenweise unterliegen Datenveröffentlichung einem zeitlich begrenzten Moratorium bis zur Veröffentlichung der dazu gehörigen Publikation."*

Gerade die eScience-Projekte beklagen jedoch, dass ihre Quellen nur eingeschränkt zugänglich sind (Hyperimage, SYNERGIE). Fehlende Standards bei den Metadaten oder Datenstrukturen können die Möglichkeiten des Austauschs zusätzlich einschränken.

Im Projekt eSciDoc, als ein Projekt, das viele Wissenschaftsdisziplinen überspannt, ist der Zugang zu Daten ein wichtiger, aber auch sensibler Punkt:

> *„eSciDoc erlaubt die gesamte Bandbreite der Zugangsbeschränkung von ‚privat' bis ‚offen'. Im Laufe des Prozesses von der ersten Idee über die Bearbeitung und Diskussion, bis zur Veröffentlichung ändern sich die Bedürfnisse an Zugriffsschutz auf die Daten. Gerade in frühen Arbeitsphasen handelt es sich teilweise um sehr sensible Daten. Die Publikation von Daten unabhängig von deren Interpretation in der Fachliteratur wird als fragwürdig angesehen. Die Veröffentlichung unbearbeiteter Rohdaten ist nicht zwingend, die Entscheidung verbleibt stets beim Wissenschaftler bzw. dem Institut."*

Der Zugang zu Daten verlangt auch, dass die Datenressourcen eindeutig identifiziert werden können. Hier kann auf die Ergebnisse des DFG-Projekts „Publikation und Zitierbarkeit wissenschaftlicher Primärdaten" (STD-DOI) zurückgegriffen werden. Darüber hinaus sind jedoch noch Verfahren zu entwickeln, die es erlauben, eindeutig auf Teilmengen sehr großer Datenbestände zu verweisen.

[36] Der Zugang zu Daten für Dritte ist in den befragten Projekten wie folgt geregelt: Freier Zugang (6), Zugang nach Vereinbarung (7), Nur Projektintern (1), Keine Policy (2). (Mehrfachnennungen waren möglich)

Zugang zu Daten über das Internet ist heute noch kein anerkannter Teil der wissenschaftlichen Kultur (Nature Redaktion, 2005; Nature Redaktion, 2006). Neben Akzeptanz müssen auch Anreize geschaffen werden, eigene Daten offen zugänglich zu machen.

> *„Bei vielen Wissenschaftlern bestehen vorbehalte, ihr Wissen frei zugänglich darzustellen. […] Ein Ansatzpunkt sind die bestehenden Reputationssysteme. Organisationen könnten mit ihrer Reputation den Einsatz neuer Werkzeuge unterstützen. Die Kultur und Praxis von Open Source Software Entwicklung könnte hier als Vorbild dienen. Offene Fragen bestehen auch zu Lizenzmodellen, Urheberrechten und der Umsetzung von Guter Wissenschaftlicher Praxis."*

Wichtig für das Vertrauen der Nutzer in die angebotenen Dienste ist eine flexible, aber dennoch klare Regelung der Zugriffsrechte, auch über lange Zeiträume hinweg. Die heute angewandten Verfahren zur Authentifizierung und Autorisierung von Nutzern sind jedoch nicht für die Verwendung über lange Zeiträume vorgesehen. Für die Authentifizierung und Autorisierung im Grid werden heute in erster Linie Zertifikate eingesetzt. Es fehlen jedoch konsistente Verfahren für die Vererbung von Rechten über lange Zeiträume, für die Archivierung der Zugriffsregelungen und für die Nachsignierung von digitalen Objekten, denn auch Verschlüsselungstechnologien „altern" und machen eine Nachsignierung mit besseren Schlüsseln notwendig. Technisch und rechtlich ungeklärt ist der Umgang mit „verwaisten" Datenbeständen, für die es keine zugriffsberechtigten Besitzer mehr gibt.

> *„Forschungsbedarf besteht bei der Frage der Rechtssicherheit in der digitale Langzeitarchivierung, z.B. der Nachsignierung von Objekten. Heutige Verfahren sind für die Sicherheit über lange Zeiträume nur bedingt geeignet."*

Gerade im Umgang mit schützenswerten Daten wird hier Misstrauen gegenüber den Systemadministratoren geäußert, denen heute noch nicht auf technischem Weg der Zugriff auf sensible Daten verwehrt werden kann. Vorfälle in den vergangen Jahren bestätigen dieses Misstrauen (Rath, 2007), wobei zusätzlich mit einer großen Dunkelziffer von Vorfällen zu rechnen ist, die nicht öffentlich bekannt wurden. Daraus entsteht in einigen Communities, insbesondere bei Partnern in der Industrie, ein Misstrauen gegenüber einer zentralen Instanz der Systemadministration, denn das Verhalten dieser Instanz wird als zu wenig transparent angesehen.

> *„Gerade für Klein- und mittelständische Unternehmen wäre dezentrale Datenhaltung und ein virtueller zentralisierte Zugang interessant. Es muss den Beteiligten jedoch die volle Kontrolle über ihre Daten garantiert werden können. Diese Rechte können heute noch nicht in der notwendigen feinen Granularität verwaltet werden. Als Folge werden die Möglichkeiten des Daten-Grid und des Datenaustausche noch unzureichend genutzt. Die Gründe dafür liegt in den Defiziten beim Management verteilter Organisationen, an Probleme bei der konsistenten Verwaltung von Zugriffsrechten und in einem Misstrauen gegenüber einer zentralen Instanz der Systemadministration, deren Handeln nicht transparent ist. Die Zugriffsrechte werden heute meist auf der Ebene des Contents geregelt."*

Diese Einschätzung von Einschränkungen bei der Nutzbarkeit des Daten-Grid teilen auch andere Projekte:

„Neue Lösungsansätze werden durch die Nutzung des Data Grid als Service erwartet. Offene Fragen sind hier Datenschutz – das beinhaltet auch die unerlaubte Kombination von Daten – und die Verwertbarkeit vor Gericht."

„Neue Lösungsansätze könnten für rechenintensive Aufgaben aus der Grid-Technologie kommen. Interessant ist auch die mögliche Realisierung eines einheitlichen Zugriffs (Single Sign-on). Auf der anderen Seite bestehen bei den Nutzern Vorbehalte gegenüber zentralisierten Diensten. Eine Zertifizierung vertrauenswürdiger Dienste und Archive könnte hier hilfreich sein."

Digital Rights Management, im Sinne von Verwertungsrechten oder Kopierschutz, spielten bei den befragten Projekten nur in Ausnahmefällen eine Rolle. Das Fiasko der Medienindustrie bei der Einführung und Durchsetzung von Digital Rights Management (DRM) Verfahren, sowie die Diskussion um den Offenen Zugang zu wissenschaftlichem Wissen betonen die Notwendigkeit, dass auch für wissenschaftliche Daten noch geeignete Lizenzmodelle gefunden werden müssen. Im Kontext der eScience- und Grid-Projekte steht hier im Vordergrund, dass die Lizenzmodelle einerseits die Weiterentwicklung von Diensten nicht behindern und die Standardisierung der Lizenzen es erlaubt, diese mit in die Daten oder Metadaten zu codieren, um die Lizenzen maschinenlesbar zu machen. Interessante Entwicklungen sind bei der Creative Commons[37] Initiative und ihrem Projekt Science Commons[38] zu beobachten. Suchmaschinen, wie z.B. Google, sind bereits heute in der Lage, standardisierte Lizenzen bei der Suche mit auszuwerten und als Filterkriterium einzusetzen.

2.4.1 Zugang zu Daten und Rechteverwaltung - Best Practice Beispiele

Einige der Best Practice Beispiele für den Zugang zu Daten wurden bereits unter den Best Practice Beispielen zur Archivtechnologie genannt (ICSU WDCs, ECMRWF, SDSS, CDS, NASA, NOAA, AHDS, OTA, DANS). Richtungsweisende Vorarbeiten wurden auch im DFG-Projekt „Publikation und Zitierbarkeit wissenschaftlicher Primärdaten" (STD-DOI) geleistet.

Als Best Practice Beispiel für die Verwaltung der Zugriffsrechte in Grid-Projkekten wurde das Projekt TeraGrid genannt, dessen Policy hier zusammen mit den Projekten „i Rule Oriented Data Systems" (iRODS) und „Storage Resource Broker" (SRB) kurz dargestellt werden.

TeraGrid Policy-based Storage Management

Für eine Anwendung in Bereichen, in denen der Zugriff auf Daten und Dienste klar geregelt sein muss, wird bemängelt, dass die Rechtevergabe nicht transparent gehandhabt wird. Dies führt insbesondere bei einer wirtschaftlichen Anwendung zu Misstrauen bei den poteziellen Nutzern. Das „policy-based storage management" im amerikanischen TeraGrid-Projekt[39] setzt für die Authentifizierung und Autorisierung der Nutzer Web-Zertifikate ein (Simmel, 2004). Bemerkenswert ist dabei, dass es eine klare Regelung über die Vergabe von Zertifikaten und den Umgang mit ihnen gibt. Hier wird der technische Aspekt einer Nutzerverwaltung, die Vertrauensbeziehungen zwischen Community-Mitgliedern als Grundlage für Authentifizierung nutzt (Choi et al., 2006), bereits erfolgreich umgesetzt. Dadurch kann die Interoperabilität zwischen Community-Grids deutlich erleichtert werden.

[37] Creative Commons: http://www.creativecommons.org
[38] Science Commons: http://sciencecommons.org
[39] TeraGrid: http://www.teragrid.org/

i Rule Oriented Data Systems (iRODS) und Storage Resource Broker (SRB)
Das Projekt "I Rule Oriented Data Systems" (iRODS)[40] entwickelt eine „Cyberinfrastruktur" für Datenmanagement. iRODS ist eine Middleware für die Verwaltung von Policies für den Zugriff auf Daten, die auf dem Storage Resource Broker (SRB)[41] aubaut. Beide Projekte werden am San Diego Supercomputing Center (SDSC) koordiniert. Der SRB ist eine Daten-Grid Anwendung, mit der heterogene Speicherressourcen einer Community über eine standardisierte Schnittstelle zur Verfügung stellt. iRODS baut auf dem SRB auf und erweitert ihn um eine ausgefeilte, regelbasierte Verwaltung der Zugriffsrechte und Verwaltung der Speicher- und Leseprozesse. Das iRODS-Konzept legt Wert darauf, dass die Prozesse dem Nutzer nicht verborgen sind, sondern von der Nutzergemeinschaft an deren Bedürfnisse angepasst werden können. Das Konzept wird von iRODS als sog. „glass box" beschrieben, im Gegensatz zu dem in Middleware-Anwendungen üblichen undurchsichtigen „black box"-Konzept.

2.5 Herausforderung Organisation und Nachhaltigkeit

2.5.1 Policy zur digitalen Langzeitarchivierung

Erfolgreiche Langzeitarchivierung von Forschungsdaten ist nicht nur eine technische Frage, sondern auch eine Frage der Organisation. Deshalb befragten wir die Projekte, ob sie selber, oder die Institutionen, an denen sie angesiedelt sind, eine eigene Policy zur Langzeitarchivierung von Daten haben.

Die Befragung ergab, dass nur eine Minderzahl der befragten Projekte eine Policy zur digitalen Langzeitarchivierung hat. Eine vergleichbare Zahl an Projekten ist dabei, eine Policy zur digitalen Langzeitarchivierung zu entwickeln. Die Policies orientieren sich, sofern kein gesetzlicher Rahmen gegeben ist, meist an den „Empfehlungen für eine Gute Wissenschaftliche Praxis" der DFG, oder ihrem Equivalent bei anderen Wissenschaftsorganisationen. Eine kleine Zahl an Projekten hat keine Policy zur digitalen Langzeitarchivierung und plant auch nicht, eine solche zu entwickeln, da digitale Langzeitarchivierung als eine Aufgabe außerhalb des Projektrahmens gesehen wird.

Im Projekt C3-Grid kommen unterschiedliche Policies parallel zum Einsatz:

> *„Die Policies über Langzeitarchivierung sind bei den einzelnen Datenprovidern unterschiedlich geregelt, teilweise durch Betriebskonzepte oder per Gesetz (z.B. „Gesetz über den Deutschen Wetterdienst). Für das World Data Center Climate beinhaltet die Policy die bewusste wissenschaftliche Entscheidung für eine Archivierung und die Dokumentation in Katalog. Alle anderen Daten werden nach Ablauf der Zeitmarke und Warnung an den Nutzer aus dem Archiv gelöscht."*

Auch wenn noch keine formelle Policy zu Langzeitarchivierung besteht, kann es ein Teil des Projekts sein, eine solche Policy zu entwickeln.

> *„Langzeitarchivierung digitaler Forschungsdaten gehört zu den Projektzielen in TextGrid, jedoch nicht in der aktuellen Projektphase. Die beteiligten Institutionen erarbeiten eine gemeinsame Policy zur digitalen Langzeitarchivierung."*

[40] iRODS: http://irods.sdsc.edu/
[41] SRB: http://www.sdsc.edu/srb/

2.5.2 Management Virtueller Organisationen

Aus den Antworten auf unsere Befragung wurde teilweise bereits sichtbar, dass der Schritt vom Projekt zur Infrastruktur noch vollzogen werden muss. Für die Langzeitarchivierung gilt in jedem Fall, dass sie weit über das Ende des jeweiligen Projekts hinaus geht. Aus dieser Diskrepanz zwischen der relativ kurzen Laufzeit der Projekte und dem Anspruch der Langzeitarchivierung erwachsen eine Reihe von Herausforderungen.

Sowohl eScience als auch Grid-Technologie sind noch sehr junge Entwicklungen. Mit dem Ende der ersten Projekte stellt sich hier die Herausforderung, die Projekte in eine nachhaltig betriebene Infrastruktur zu überführen. Entscheidend sind dabei die Organisationsmodelle für Virtuelle Organisationen, die eine Community-Infrastruktur als kollaborative Einrichtung betreiben. (Edwards et al., 2007) beschreiben in ihrer Studie für die US National Science Foundation (NSF) Spannungsfelder, die Quellen für Konflikte bei der Überführung einer Virtuellen Organisation von der Projektphase in eine dauerhaft betriebene Infrastruktur.

- Wie werden Zielkonflikte zwischen kurzfristiger Projektfinanzierung und einem langfristigen Aufbau einer Infrastruktur aufgelöst?
- Wer bestimmt die Richtlinien über den Zugang zu Daten und für deren Archivierung?
- Wie werden Konflikte zwischen lokalen Praktiken und globalen Standards der aufgelöst, die einer effektiven Zusammenarbeit im Weg stehen?
- Wie kann eine Community-Infrastruktur weiterentwickelt werden, ohne die Möglichkeit einer internationalen oder globalen Infrastruktur zu erschweren oder gar zu verhindern?
- Wer hat die Kompetenz, Richtlinien für die weitere Entwicklung der Infrastruktur zu bestimmen?
- Wie werden neu entstehende Technologien in die bereits bestehende Infrastruktur integriert?

Ein immer wieder genannter Punkt sind Unsicherheiten über die Vertrauenswürdigkeit von digitalen Archiven. Insbesondere in Fällen, in denen eine rechtlich begründete Pflicht zur Archivierung gegeben ist, bestehen Fragen zur Rechtssicherheit in der digitalen Langzeitarchivierung. Um digitale Objekte persistent zum Zweck der Langzeitarchivierung im Daten-Grid ablegen zu können ist es notwendig, mit den Anbietern der Speicherkapazität im Daten-Grid entsprechende Service Level Agreements zu vereinbaren. Hier sollten die Maßstäbe des Kriterienkatalogs vertrauenswürdige digitale Langzeitarchive angelegt werden (Dobratz et al., 2006).

In diesem Zusammenhang wird auch das Management von Virtuellen Organisationen (VO) in Frage gestellt, ob dieses auf lange Zeiträume hin betrachtet belastbar ist.

> *„Forschungsbedarf besteht auch noch für das Management von VO Strukturen. In der Praxis sind oft die Zuständigkeiten in einer VO nicht geklärt.“*

Bemerkenswert ist an dieser Stelle, dass die Abrechnung von Leistungen im Grid innerhalb von VO und mit Dritten nicht als noch offene Frage wahrgenommen wird, denn aus dem Kreis der befragten Projekte kamen hierzu keine Aussagen Im Rahmen zeitlich befristeter Community-Projekte mag sich die Frage nicht in dieser Form stellen. Über das Ende des Projektes hinaus entstehen jedoch weiterhin reale Kosten durch die digitale Langzeitarchivierung der Daten, denn Medien- oder Formatmigration können erhebliche zusätzliche Kosten erzeugen. Für die Abrechnung von Leistungen in der digitalen Langzeitarchivierung besteht noch kein Modell der Finanzierung, wenngleich es bereits

Ansätze gibt, Grid-Leistungen und Aufwendungen zur Langzeitarchivierung zwischen Nutzern und Anbietern abzurechnen (siehe z.B. Abdelkader und Broeckhove, 2007; Lavoie, 2003).

Da die Projekte sich selber nicht als permanente Infrastruktur sehen, sondern als Pilotprojekte zum Aufbau einer Infrastruktur stellt sich für die Projekte die Frage, wie Anreize für die Wissenschaftler geschaffen werden können, die neu aufgebauten Systeme auch zu nutzen. Mit dieser Frage beschäftigen sich bereits Projekte im Kontext der britische eScience-Förderung (Lavoie, 2003). Keine der Infrastrukturen für eine digitale Langzeitarchivierung lässt sich dauerhaft betreiben, wenn es keine Nutzer gibt. Erst wenn eine Nachfrage der Wissenschaft nach einer digitalen Langzeitarchivierung besteht, können dauerhafte Strukturen entstehen. Aus diesem Grund ist es entscheidend, Anreize zur digitalen Langzeitarchivierung von Forschungsdaten zu schaffen.

> *„Gerade die Langzeitarchivierung wird als Anreiz für die Wissenschaftler gesehen, ihre Quellen und Werkzeuge TextGrid zur Verfügung zu stellen."*

> *„Es müssen daher Anreize gefunden werden, mit denen Wissenschaftler dazu bewogen werden können, die angebotenen Werkzeuge auch zu nutzen. Ein Ansatzpunkt sind die bestehenden Reputationssysteme. Organisationen könnten mit ihrer Reputation den Einsatz neuer Werkzeuge unterstützen. Die Kultur und Praxis von Open Source Software Entwicklung könnte hier als Vorbild dienen."*

Die Erfahrungen bei der Umsetzung von Open Access Policies haben gezeigt, dass Selbstverpflichtungen wirkungslos sind (Zerhouni, 2006). Auch Verpflichtungen durch die Forschungseinrichtungen oder Förderorganisationen sind nur erfolgreich, wenn sie neben Anreizen auch mit Sanktionen bewehrt sind (Spittler, 1967). Im Gegenzug muss den Wissenschaftlern entsprechende technische Unterstützung zur digitalen Langzeitarchivierung angeboten werden (Lyon, 2007).

Offen ist auch, wer die Kompetenz zur digitalen Langzeitarchivierung trägt und wie diese finanziert wird. Ähnlich wie in der Frage der digitalen Langzeitarchivierung von Forschungsdaten ist auch bei den eScience- und Grid-Projekten oft die Frage, wie eine digitale Langzeitarchivierung der Ergebnisse der Projekte durchgeführt werden soll, jenseits den Planungshorizonts. Den Projekten sollte deshalb eine professionelle Unterstützung bei der Planung und Durchführung der digitalen Langzeitarchivierung angeboten werden. Diese Kompetenz muss gezielt aufgebaut werden. In den durchgeführten Interviews wurde auch deutlich, dass kaum Best-Practice Beispiele bekannt sind, in denen vorbildliche Lösungen für Anforderungen von Grid-Technologie und eScience an die digitale Langzeitarchivierung von Forschungsdaten gefunden wurden.

> *„Oft sind vorhandene Technologien nicht bekannt oder in der Praxis noch nicht akzeptiert. Deshalb sollte der Technologietransfer zwischen Projekten gefördert werden."*

3 Handlungsempfehlungen

Im vorangegangenen Abschnitt wurden die Herausforderungen an die digitale Langzeitarchivierung in eScience- und Grid-Projekten identifiziert. Aus diesen Herausforderungen werden in diesem Abschnitt Handlungsempfehlungen abgeleitet, die sich an Betreiber von Community-Grids, Nutzer von Community-Grids und an Entscheidungsträger in den Fachgremien und Lenkungsausschüssen im Umfeld der Community-Grids richten.

Besonderes Augenmerk richtet diese Studie auf die Nachnutzbarkeit von Forschungsdaten. Diese können aus experimentellen Aufbauten stammen, aus Monitoringsystemen, aber auch aus der Geistes- und Sozialwissenschaftlichen Forschung. Um nachnutzbar zu sein, müssen diese Daten in digitalen Langzeitarchiven vorliegen und dort findbar sein. Die Metadaten müssen dem Nutzer nicht nur den Inhalt der Daten darstellen, sondern auch den Kontext, aus dem heraus sie erzeugt wurden, damit der Nutzer ihre Nachnutzbarkeit und Vertrauenswürdigkeit abschätzen kann.

3.1 Technik

Zum jetzigen Stand ist es noch nicht klar, welche Grid-Technologien in welchen Prozessen des OAIS-Refenzmodells eingesetzt werden können. In einem Testbed könnten anwendungsnah Grid-Dienste erprobt werden, die für digitale Langzeitarchivierung benötigt werden und den Transfer von Grid-Technologie in die digitale Langzeitarchivierung zu unterstützen. Aus diesem Testbed heraus könnte ein Community-Grid für digitale Langzeitarchivierung entwickelt werden.

Da die Entwicklung von Standards in der Grid-Technologie nicht von einer zentralen Stelle koordiniert wird, beklagen Anwender außerhalb der Community-Grids, dass die aktuellen Standards noch nicht stabil genug sind. Ein Testbed für den Einsatz von Grid-Diensten für die digitale Langzeitarchivierung könnte dazu beitragen, die hier relevanten Standards zu stabilisieren (Schiffmann, in prep.). Weitere Handlungsempfehlungen hierzu werden in der entsprechenden nestor-Expertise (Borghoff und Rödig, in prep.) formuliert.

Forschungsbedarf besteht in der Frage, wie Daten genutzt werden können, wenn die Plattformen, auf denen sie erzeugt, bearbeitet und visualisiert wurden, nicht mehr vorhanden sind, weil sie als veraltet ausgemustert wurden. Grid-basierte Anwendungen könnten bei der Transformation in neue Formate und Langzeitarchivierungsumgebungen eingesetzt werden, als auch bei der Emulation technisch obsoleter Betriebssystemumgebungen.

Ebenso wie Abspielumgebungen werden auch Dateiformate mit der Zeit obsolet und erfordern eine Formatmigration. Nicht alle Dateiformate, selbst wenn sie in der Praxis weite Verbreitung finden, sind offen dokumentiert und nicht alle lassen sich verlustfrei migrieren. Ein Kriterienkatalog für die Archivfähigkeit von Dateiformaten würde helfen die Archivierungsstrategien zu verbessern.

Auffallend, nicht nur in der Auswertung der Befragung, ist der geringe Bekanntheitsgrad von Best-Practice Beispielen. Die Vermittlung von Best-Practice Beispielen aus eScience- und Grid Projekten wäre eine geeignete Maßnahme, für die Vermittlung von Konzepten und Kompetenzen für die digitalen Langzeitarchivierung von Forschungsdaten und zur Verbesserung der aktuellen Situation.

3.2 Metadaten

Angemessene Metadaten sind eine Schlüsselkomponente für die Interoperabilität von Diensten. Ebenso sind Metadaten notwendig, um Datenbestände in Katalogen und Fachportalen nachzuweisen und findbar zu machen. In einigen Communities besteht noch Bedarf bei der Formulierung von Metadatenstandards und Anwendungsprofilen. Auch hier sind Best-Practice Beispiele in den Projekten wenig bekannt. Veranstaltungen wie der D-GRID Metadaten-Workshop im März 2007 an der SUB Göttingen unterstützen die Standardisierungsprozesse durch die Vermittlung von Best-Practice Beispielen.

In Grid-Projekten bestehen gute Voraussetzungen für die Entwicklung von Verfahren, mit denen die Beschreibung von Forschungsdaten durch Metadaten unterstützt werden kann. Diese Verfahren können in Workflows in der Prozessierung und in der Redaktion von Datenobjekten eingesetzt werden, in einigen Fällen wäre sogar eine automatische Annotation möglich. Dabei sollten Metadaten, welche die Herkunft und Bearbeitung der Daten beschreiben, verstärkt beachtet werden, da die Kodifizierung des Prozesswissens die Möglichkeiten der Nachnutzung von wissenschaftlichen Primärdaten entscheidend verbessern.

3.3 Semantic Grid

Aufbauend auf Arbeiten zur Verbesserung des Umgangs mit Metadaten sollte in Grid-Projekten Verfahren des Semantic Web aus eScience-Projekten übertragen werden. Die verbesserte semantische Interoperabilität, auch Semantic Grid genannt, würde den Austausch von Daten und Diensten zwischen Community-Grids erleichtern und damit die Nachnutzbarkeit archivierter Daten verbessern.

Für den Aufbau eines Semantic Grid ist es notwendig, dass Datenressourcen eindeutig identifiziert werden können. Bereits bestehende Verfahren, die auf eindeutigen, global auflösbaren Identifikatoren beruhen, müssen noch erweitert werden um auch kleine Teilmengen aus sehr großen Datenbeständen eindeutig referenzieren zu können. Die Findbarkeit dieser Daten könnte zusätzlich durch die Entwicklung von effizienten Vorschauverfahren für große, mehrdimensionale Datensätze erhöht werden.

3.4 Rechteverwaltung

Für den Transfer von Grid-Technologie in die digitale Langzeitarchivierung von Forschungsdaten sind im Bereich der Verwaltung von Zugriffsrechten noch einige sektorspezifische Herausforderungen zu überwinden, z.B. der Schutz von Persönlichkeitsrechten (z.B. Medizin) oder vertraulichen Unterlagen (wirtschaftliche Anwendungen). Insbesondere müssen die heute verfügbaren Authentifizierungs- und Autorisierungsverfahren im Grid darauf hin geprüft werden, ob sie sich für den Einsatz über lange Zeit eignen und ob die Regelungen bei einem Technologiewandel auf neue Verfahren sicher übertragen werden können.

Für verteilte Systeme und Serviceorientierte Architekturen werden auch neue Konzepte für Authentifizierung und Autorisierung unter dem Namen „Identity 2.0“ diskutiert (Choi et al., 2006). Die hier diskutierten Ansätze von Identity Credential Services kommen der verteilten Struktur des Grid sehr entgegen und könnten in Zukunft zur Rechteverwaltung in Grid-Diensten beitragen (Simmel, 2004).

3.5 Management Virtueller Organisationen

Um die Ergebnisse der eScience- und Grid-Projekte erfolgreich in eine Grid-Infrastruktur zu überführen, die von neuen eScience- und Grid-Projekten nachgenutzt wird, müssen Mangementmodelle für Virtuelle Organisationen entwickelt werden. Insbesondere fehlen Management- und Kostenmodelle für die digitale Langzeitarchivierung als Community-Grid oder als Grid-Dienst, wie auch umgekehrt für die Nutzung von Grid-Diensten für die digitale Langzeitarchivierung.

Um die Praxis der digitalen Langzeitarchivierung von Forschungsdaten zu verbessern sollte auch untersucht werden, worin für die Akteure im Umfeld von Forschungsdaten die Anreize zur digitalen Langzeitarchivierung bestehen, oder wie Anreize geschaffen werden können (Lavoie, 2003) und bei wem die Verantwortung für die Langzeiterhaltung digitaler Forschungsdaten liegt (Lyon, 2007). Eine entsprechende Management- und Förderpraxis, die mehr Wissenschaftler an der Entwicklung der Infrastruktur beteiligt, würde zu eine nachhaltigen Entwicklung beitragen. Dabei sollte auch eine Zusammenarbeit mit Organisationen angestrebt werden, die über Expertise komplementär zur Forschung in den Community Grids verfügen, wie z.B. Archive oder sozialwissenschaftliche Arbeisgruppen.

Für die Weiterentwicklung der digitalen Langzeitarchivierung von Forschungsdaten aus eScience- und Grid-Projekten wurden bereits Maßnahmen zur Vernetzung der Projekte untereinander ergriffen. In Zukunft ist auch eine Professionalisierung notwendig, um die aufgebauten Infrastrukturen langfristig betreiben zu können. Auch hier kann, wie bei den technischen Herausforderungen, von Best-Practice Beispielen in anderen Bereichen und in anderen Ländern gelernt werden.

Danksagung

Diese Expertise wurde im Rahmen des Projekts „Kompetenznetzwerk Langzeitarchivierung“ (nestor) erstellt, das vom Bundesministerium für Forschung und Bildung gefördert wird.

Als Autor möchte ich mich bei meinen Gesprächspartnern in den eScience- und Grid-Projekten für ihre Bereitschaft bedanken, die Expertise mit ihrem Wissen und ihrer Zeit zu unterstützen. Desgleichen gilt mein Dank den Mitgliedern der nestor-AG „Grid/eScience“ und den Teilnehmern am nestor-Workshop im Rahmen der German eScience Conference (GES2007) in Baden-Baden im Mai 2007 für ihre Beiträge zur Diskussion, die bei der Erstellung der Expertise sehr hilfreich waren.

Literatur

Abdelkader, K. und Broeckhove, J., 2007. Pricing Resources in Dynamic Grid Economies. In: W. Bühler (Hrsg.), German e-Science Conference. Max Planck Digital Library, Baden-Baden.

Barga, R. und Gannon, D.B., 2007. Scientific versus business workflows. In: I.J. Taylor, E. Deelman, D.B. Gannon und M. Shields (Hrsg.), Workflows for e-Science. Springer-Verlag, London, Großbritannien, S. 9-16.

Berliner Erklärung, 2003. Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, Berlin, S. 3.

Berman, F., Fox, G.C. und Hey, T., 2003. The Grid as the Future Computing Infrastructure. In: F. Berman, G.C. Fox und T. Hey (Hrsg.), Grid Computing. Wiley InterScience, Hoboken, NJ, S. 9-50.

Berman, H., Henrick, K., Nakamura, H. und Markley, J.L., 2007. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Research, 35(suppl. 1): D301-303. doi:10.1093/nar/gkl971

Borghoff, U.M. und Rödig, P., in prep. Standards und Standardisierung im Kontext von Grid-Technologien und Langzeitarchivierung. nestor-Materialien, Kompetenznetzwerk Langzeitarchivierung (nestor), Göttingen.

Brase, J., 2004. Using Digital Library Techniques - Registration of Scientific Primary Data. In: M. Jones, E.A. Fox und R. Shen (Hrsg.), Research and Advanced Technology for Digital Libraries. Lecture Notes in Computer Science. Springer-Verlag, Heidelberg, S. 488-494.

Choi, H.-C. et al., 2006. Trust Models for Community Aware Identity Management, WWW2006, Edinburgh, Großbritannien.

Cotter, G., Frame, M. und Sepic, R., 2004. Integrated science for environmental decision-making: the challenge for biodiversity and ecosystem informatics. Data Science Journal, 3: 38-59. doi:10.2481/dsj.3.38

Curtis, J., Koerbin, P., Raftos, P., Berriman, D. und Hunter, J., 2007. AONS - An obsolescence detection and notification service for Web archives and digital repositories. New Review of Hypermedia and Multimedia, 13(1): 39-53. doi:10.1080/13614560701423711

De Roure, D., Jennings, N.R. und Shadbold, N.R., 2005. The Semantic Grid: Past, Present and Future. Proceedings of the IEEE, 93(3): 669-681. http://eprints.ecs.soton.ac.uk/9976/

DFG, 1998. Regeln guter wissenschaftlicher Praxis, Deutsche Forschungsgemeinschaft. http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/self_regulation_98.pdf

Dobratz, S. et al., 2006. Kriterienkatalog vertrauenswürdige digitale Langzeitarchive. nestor Materialien, 8 (Version 1), Die Deutsche Bibliothek, Frankfurt (Main). urn:nbn:de:0008-2006060710, http://edoc.hu-berlin.de/series/nestor-materialien/2006-8/PDF/8.pdf

Eastman, T.E. et al., 2005. eScience and archiving for space science. Data Science Journal, 4: 67-76. doi:10.2481/dsj.4.67

Edwards, P.N., Jackson, S.J., Bowker, G.C. und Knobel, C.P., 2007. Understanding Infrastructure: Dynamics, Tensions, and Design, National Science Foundation, Washington, D.C., USA. http://hdl.handle.net/2027.42/49353

Fornwall, M., 2004. Relationship between OBIS and other national and international biodiversity information systems, USGS. Reston, VA, USA.

Genova, F. et al., 2005. Running a data centre on the long-term: Lessons learnt from 30 years of CDS history, Ensuring Long-term Preservation and Adding Value to Scientific and Technical data (PV 2005). UKOLN, Edinburgh, Großbritannien.

Hey, T. und Trefethen, A., 2003a. The data deluge: an eScience perspective. In: F. Berman, T. Hey und G.C. Fox (Hrsg.), Grid Computing - Making the Global Infrastructure Reality. Wiley & Sons, Ltd., New York, NY, USA, S. 409-435.

Hey, T. und Trefethen, A., 2003b. e-Science and its implications. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 361(1809): 1809-1825. doi:10.1098/rsta.2003.1224

Hitchcock, S., Brody, T., Hey, J.M.N. und Carr, L., 2007. Digital Preservation Service Provider Models for Institutional Repositories. D-Lib Magazine, 13(5/6): 16. doi:10.1045/may2007-hitchcock

Huber, R. und Schindler, U., 2007. Open Geo-Archives - Integrating earthscience data centers into research portals, European GeoInformatics Workshop. e-Science Institute, Edinburgh, Großbritannien, S. 10.

Kindermann, S., Stockhause, M. und Ronneberger, K., 2006. Intelligent Data Networking for the Earth System Science Community. In: W. Bühler (Hrsg.), German eScience Conference. Max Planck Digital Library, Baden-Baden, S. 10.

Klump, J. et al., 2006. Data publication in the Open Access Initiative. Data Science Journal, 5: 79-83. doi:10.2481/dsj.5.79

Klump, J., Löwe, P., Häner, R. und Wächter, J., 2007. Continuous digital workflows for earth science research. In: W. Bühler (Hrsg.), German eScience Conference. Max Planck Digital Library, Baden-Baden, S. 8.

Kroker, H., 2006. Wissenschaftlicher Fortschritt hängt immer stärker von der Verarbeitung gewaltiger Datenmengen ab. Die Welt. 2006-03-29 http://www.welt.de/data/2006/03/29/866493.html.

Lavoie, B., 2003. The Incentives to Preserve Digital Materials: Roles, Scenarios, and Economic Decision-Making, OCLC Online Computer Library Center, Inc., Dublin, OH, USA. http://www.oclc.org/research/projects/digipres/incentives-dp.pdf

Lord, P. und Macdonald, A., 2003. e-Science Curation Report - Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision, JISC. http://www.jisc.ac.uk/uploaded_documents/e-scienceReportFinal.pdf

Lormant, N., Huc, C., Boucon, D. und Miquel, C., 2005. How to Evaluate the Ability of a File Format to Ensure Long-Term Preservation for Digital Information?, Ensuring Long-term Preservation and Adding Value to Scientific and Technical data (PV 2005), Edinburgh, Großbritannien, S. 11.

Lyon, L., 2007. Dealing with Data: Roles, Rights, Responsibilities and Relationships, UKOLN, Bath, Großbritannien. http://www.jisc.ac.uk/media/documents/programmes/digital_repositories/dealing_with_data_report-final.pdf

Nature Redaktion, 2005. Let data speak to data. Nature, 438(7068): 531. doi:10.1038/438531a

Nature Redaktion, 2006. A fair share. Nature, 444(7120): 653-654. doi:10.1038/444653b

OAIS, 2002. Reference Model for an Open Archival Information System (OAIS). Blue Book., CCSDS 650.0-B-1, Consultative Committee for Space Data Systems, Greenbelt, MD, USA. http://public.ccsds.org/publications/archive/650x0b1.pdf

OECD, 2004. Science, Technology and Innovation for the 21st Century. Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004 - Final Communiqué, Organisation for Economic Co-operation and Development, Paris, Frankreich. http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html

OECD, 2006. Recommendation of the Council concerning Access to Research Data from Public Funding, C(2006)184, Organisation for Economic Co-operation and Development, Paris, Frankreich. http://webdomino1.oecd.org/horizontal/oecdacts.nsf/Display/3A5FB1397B5ADFB7C12572980053C9D3?OpenDocument

Rath, C., 2007. Angst vor indiskreten Telekom-Beschäftigten - Linksliberale Richter gegen Vorratsspeicherung von Verbindungsdaten. Hacker und Spione könnten Zugriff bekommen. die tageszeitung: 7. 2007-07-18 http://www.taz.de/index.php?id=digi-artikel&ressort=in&dig=2007/07/18/a0106&menu=1.

Rothenberg, J., 1997. Digital Information Lasts Forever—Or Five Years, Whichever Comes First. RAND Corp.

Schiffmann, W., in prep. Synergiepotenziale zwischen GRID- und e-Science-Technologien für die Langzeitarchivierung. nestor-Materialien, Kompetenznetzwerk Langzeitarchivierung (nestor), Göttingen.

Schindler, U., Bräuer, B. und Diepenbroek, M., 2007. Data Information Service based on Open Archives Initiative Protocols and Apache Lucene. In: W. Bühler (Hrsg.), German eScience Conference. Max Planck Digital Library, Baden-Baden.

Schroeder, R., den Besten, M. und Fry, J., 2007. Catching Up or Latecomer Advantage? Lessons from e-Research Strategies in Germany, in the UK and Beyond. In: W. Bühler (Hrsg.), German eScience Conference. Max-Planck Digital Library, Baden-Baden, S. 8.

Severiens, T. und Hilf, E.R., 2006a. Studie zum Stand vorhandener Forschungsdaten und Rohdaten aus wissenschaftlichen Tätigkeiten: Erfordernisse und Eignung zur Archivierung bzw. Zurverfügungstellung in Deutschland (Primärdaten). nestor Materialien, 6, nestor - Kompetenznetzwerk Langzeitarchivierung, Göttingen. urn:nbn:de:0008-20051114018

Severiens, T. und Hilf, E.R., 2006b. Zur Entwicklung eines Beschreibungsprofils für eine nationale Langzeit-Archivierungs-Strategie - ein Beitrag aus der Sicht der Wissenschaften. nestor Materialien, 7, nestor - Kompetenznetzwerk Langzeitarchivierung, Göttingen. urn:nbn:de:0008-20051114018

Simmel, D., 2004. TeraGrid Certificate Management and Authorization Policy, Pittsburgh Supercomputing Center, Carnegie Mellon University, University of Pittsburgh, Pittsburg, PA, USA. http://www.teragrid.org/policy/TGCertPolicy-TG-5.pdf

Spittler, G., 1967. Norm und Sanktion. Untersuchungen zum Sanktionsmechanismus. Walter Verlag, Olten, Schweiz, 153 S.

Uhlir, P.F. und Schröder, P., 2007. Open Data for Global Science. Data Science Journal, 6(Open Data Issue): OD36-53. doi:10.2481/dsj.6.OD36

Zerhouni, E.A., 2006. Report on the NIH public access policy, National Institute of Health, Bethesda, MD, USA. http://publicaccess.nih.gov/Final_Report_20060201.pdf

Anhang – Fragebogen

„Anforderungen von eScience und Grid-Technologie an die Archivierung wissenschaftlicher Rohdaten“

In der wissenschaftlichen Forschung produzierten Daten, zum Beispiel aus Messungen oder Experimenten, sind in vielen Sektoren von zentraler Bedeutung. Sowohl öffentliche Institutionen wie auch kommerzielle Unternehmen investieren erhebliche Mittel in die Produktion von Rohdaten und das jährlich produzierte Volumen an Rohdaten steigt stetig an. Damit gewinnt auch die Forderung nach deren Verfügbarkeit zur möglichen Nachprüfung von wissenschaftlichen Ergebnissen und zur Wiederverwendung große Bedeutung.

Gerade wegen dieser extremen Anforderungen an Speicherressourcen und zusätzlichen Managementvorkehrungen sind die Rohdatenerzeugenden Communities in der Anwendung von Grid-Technologien vergleichsweise weit fortgeschritten. Astrophysik, Klimatologie, biomedizinische Forschung, und andere Communities wenden bereits seit einiger Zeit Grid-Technologien an.

Die Expertise „Anforderungen von eScience und Grid-Technologie an die Archivierung wissenschaftlicher Rohdaten“ soll sowohl aus technologischer wie organisatorisch-strategischer Perspektive prüfen, ob existierende e-Science-Infrastrukturen in Rohdatenproduzierenden Communities den Anforderungen zur Langzeitarchivierung (länger als den in der Community üblichen Aufbewahrungsdauern von etwa 10 Jahren) gerecht werden können, und ob die Erfahrungen der Communities im Bereich der Grid-Technologien auf Organisationen und Systeme zur Langzeitarchivierung übertragen werden können.

1 Erwartete Daten: Menge und Komplexität

Grid-Anwendungen zeichnen sich durch sehr große Datenmengen aus. In eScience-Projekten sind die erwarteten Datenmengen im Vergleich dazu wesentlich kleiner, sind jedoch durch eine hohe semantische Komplexität gekennzeichnet.

1. Welche Datenmengen erwarten Sie in Ihrem Projekt? (Datenmenge, Anzahl der Objekte)
2. Welchen Grad semantischer Komplexität Ihrer Daten erwarten Sie?

Die DFG, und andere Wissenschaftsorganisationen, empfehlen als gute Wissenschaftliche Praxis die Archivierung von Daten für einen Zeitraum von mindestens zehn Jahren.

3. Für wie lange sollen die Daten aus Ihrem Projekt archiviert werden?

Nicht alle Daten müssen der Nachwelt erhalten bleiben. Insbesondere bei sehr großen Datenmengen muss eine Auswahl getroffen werden.

4. Werden alle Daten archiviert? Welche Auswahlkriterien gibt es?

Strategien der Langzeitarchivierung müssen sich auch mit Daten- und Medientypen auseinandersetzen, denn nicht jeder Daten- und Medientyp eignet sich gleichermaßen für die Langzeitarchivierung digitaler Objekte.

5. Welche Daten- und Medientypen erwarten Sie in Ihrem Projekt?

2 Umgang mit Metadaten

Reine Datendateien sind oft ohne Beschreibung ihrer Struktur, ihrer Herkunft oder ihrer Benutzung schon nach kurzer Zeit nicht mehr nutzbar. Disziplin-spezifische Beschreibungen der Daten helfen, diese zu lokalisieren und nach zu nutzen. Aus diesem Grund messen wir Metadaten eine hohe Bedeutung bei.

6. Welche Metadaten werden in Ihrem Projekt gespeichert?
    a. Speicherung und Zugang zu den Daten
    b. Herkunft und Prozessierung der Daten
    c. Benutzung (technisch) der Daten
    d. Beschreibende Disziplin-spezifische Metadaten

In der Praxis hat es sich bewährt, wenn die Beschreibung der Metadaten einem anerkannten Standard folgt, damit die Bedeutung der beschreibenden Attribute dokumentiert und möglichst eindeutig ist und auch zu einem späteren Zeitpunkt noch verstanden werden kann.

7. Folgen die Metadaten-Profile in Ihrem Projekt anerkannten Standards?

Insbesondere für die Nutzung von Daten mit Werkzeigen des eScience ist es interessant, wenn auch die Beziehungen zwischen Datensätzen aufgezeichnet und archiviert werden.

8. Wie werden semantische Beziehungen zwischen Daten repräsentiert?

In die Erstellung der Daten fließt einiges an implizitem Prozesswissen, das über Herkunft und Prozessierung der Daten hinaus geht.

9. Wird mit den Daten auch Prozesswissen archiviert?

3 Daten-Grid und digitale Bibliotheken

In der Diskussion über digitale Bibliotheken und den offenen Zugang zu wissenschaftlichem Wissen ist auch der Zugang zu Daten und deren Austausch unter Wissenschaftlern Gegenstand des Diskurses über die Zukunft der Forschung geworden.

10. Sind die Daten für Dritte zugänglich? (Data sharing)

Forscher wollen vorhandene Daten nachnutzen. Um Inhalt, Qualität und damit die Nachnutzbarkeit der Daten einschätzen zu können, orientieren sie sich an der Interpretation der Daten in der wissenschaftlichen Fachliteratur. Zudem gibt es Fälle, z.B. in der Biologie oder Geologie, in der es für die Forscher interessant ist, das Objekt, an dem die Daten erhoben wurden, identifizieren zu können.

11. Werden in Ihrem Projekt semantische Verbindungen zwischen Veröffentlichungen, Daten und Forschungsmaterialien mit verwaltet? (Semantic Web und Internet der Dinge)

4 Forschungsbedarf

Mit den großen Datenmengen der Grid-Projekte und den semantisch komplexen Daten der eScience-Projekte kommen neue Herausforderungen auf die Langzeitarchivierung von Forschungsdaten zu.

12. Wo sehen sie Forschungsbedarf in Bezug auf neue Anforderungen an die Archivierung wissenschaftlicher Rohdaten durch das Aufkommen von eScience und Grid-Technologie?

Die Service-Orientierte Architektur der Grid-Technologie biete für die Langzeitarchivierung (LZA) potenziell neue Lösungsansätze, in dem Rechen- oder Speicherintensive Prozesse aus LZA-Anwendungen in Grid-Prozesse ausgelagert werden.

13. Erwarten Sie durch Grid-Technologie neue Lösungsansätze für die Langzeitarchivierung wissenschaftlicher Primärdaten? (Data Grid, Archive Ingest/Extraction, Media/Format Migration, Autorisierung)

5 Best Practice

Erfolgreiche Langzeitarchivierung von Forschungsdaten ist nicht nur eine technische Frage, sondern auch eine Frage der Organisation.

14. Hat Ihre Einrichtung eine Policy über die Langzeitarchivierung von Daten?

Bei der Umsetzung von der Theorie in die Praxis werden die Stärken und Schwächen einer Policy zur Langzeitarchivierung von Forschungsdaten sichtbar. In einigen Projekten wurden bereits Erfahrungen gesammelt.

15. An welcher Stelle sehen sie die Anforderungen von eScience und Grid-Technologie an die Langzeitarchivierung digitaler Forschungsdaten vorbildlich gelöst?